Prediction of Writing True Scores in Automated Scoring of Essays by Best Linear Predictors and Penalized Best Linear Predictors

ETS RR–19-13

Lili Yao
Shelby J. Haberman
Mo Zhang

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Prediction of Writing True Scores in Automated Scoring of Essays by Best Linear Predictors and Penalized Best Linear Predictors

Lili Yao,[1] Shelby J. Haberman,[2] & Mo Zhang[1]

1 Educational Testing Service Princeton, NJ
2 Consultant, Jerusalem, Israel

Many assessments of writing proficiency that aid in making high-stakes decisions consist of several essay tasks evaluated by a combination of human holistic scores and computer-generated scores for essay features such as the rate of grammatical errors per word. Under typical conditions, a summary writing score is provided by a linear combination of the holistic scores and the feature scores. The best linear predictor (BLP) is used to approximate the true composite writing score by a linear combination of holistic scores and scores of essay features. However, because the relationship between computer-generated feature score and human scores may depend on subgroup membership and the same scoring rules must normally be applied to all test takers, Yao, Haberman, and Zhang proposed a modified methodology of the penalized best linear predictor (PBLP) by incorporating a quadratic penalty function into the conventional BLP method. This research report contains full accounts of the BLP results as well as supplementary PBLP results to Yao et al. for three assessments of writing that aid in making high-stakes decisions: the *TOEFL iBT*® Writing test, the *GRE*® General Analytical Writing subject test, and the *Praxis*® Core Academic Skills for Educators: Writing assessment. Results obtained indicate the added value in using machine features for prediction of composite true scores of essay writings and effectiveness of the penalty function in suppressing the lack of population invariance.

Keywords Automated scoring; writing assessment; best linear predictor; penalty function

doi:10.1002/ets2.12248

Classical test theory (Lord & Novick, 1968) may be applied to develop scores for writing assessments that consist of several essay tasks for which both human holistic scores and computer-generated scores of essay features are available. Under typical conditions, the summary writing score is based on a linear combination of the holistic scores and the feature scores. The linear combination is selected to predict a true composite writing score that corresponds to an observed weighted average of human holistic scores. Best linear prediction (BLP; Haberman, 2008; Haberman & Qian, 2007; Haberman & Yao, 2015; Haberman, Yao, & Sinharay, 2015; Wainer, Sheehan, & Wang, 2000; Wainer et al., 2001) may be used to find the best linear combination of observed human scores and computer-generated feature scores for prediction of the true composite writing score, to minimize the mean squared error that is the expected square of the residual difference between the true composite score and its linear prediction. In the case of scoring accuracy, the true composite writing score is the expected composite writing score achieved by random selection from a population of human raters for the two specific essays observed. In the case of assessment accuracy, the true composite writing score is the expected composite writing score obtained by a test taker tested on a randomly selected test form from a population of parallel writing assessments. For example, the *TOEFL iBT*® Writing test (*TOEFL*® Writing) consists of two essay prompts, an integrated prompt and an independent prompt. Human raters use scoring rubrics to provide an integer holistic score for each prompt. Normally, for each of the two prompts, the *e-rater*® automated scoring engine (Attali & Burstein, 2006; Attali, Burstein, & Andreyev, 2003; Burstein, Chodorow, & Leacock, 2004) provides numerical measures (feature scores) of aspects of essay quality, such as grammatical correctness and syntactic structure. A possible observed composite score is the arithmetic average of the observed holistic score on the integrated prompt and the observed holistic score on the independent prompt. For either scoring accuracy or assessment accuracy, the true composite score corresponds to this average. With BLP, the true

Corresponding author: L. Yao, E-mail: lyao@ets.org

composite score is then estimated from both the observed holistic scores on the two prompts and from the observed essay feature scores from the two prompts.

When BLP is used to predict the true composite score from both human holistic scores and computer-generated essay features, problems of fairness can arise if the relationship of human holistic scores to computer-generated essay features is not the same for all subgroups of interest. For example, in TOEFL Writing, fairness concerns arise if the relationship of human holistic scores to computer-generated essay features for Chinese test takers is different from the corresponding relationship for Japanese test takers. Given the general principle that two test takers with the exact response quality should receive the same test score, this issue can result in scores based on BLP rather than on observed composite scores that have a differential impact on Chinese test takers relative to Japanese test takers. This problem is closely related to the problem of population invariance in equating (Dorans & Holland, 2000) and to the problem of subgroups in score augmentation (Haberman & Sinharay, 2013). To treat the issue of subgroup biases, Yao, Haberman, and Zhang (2019) developed a generalized version of BLP as the penalized best linear predictor (PBLP). With PBLP, linear predictors are found by minimizing the sum of mean squared error of prediction and a quadratic penalty function. If the penalty function is always 0, then PBLP reduces to BLP. In general, the penalty function relies on division of the population of test takers into subgroups defined in terms of variables like testing country, native language, or race/ethnicity. The quadratic penalty function is a nonnegative multiple of the mean squared conditional expected residual difference between the true composite score and its linear prediction given subgroup membership. The value of the modification of BLP is assessed via examination of two generally competing criteria: the mean squared error of prediction of the true composite score and the mean difference in subgroup means for observed composite score and rescaled predicted score, where the rescaled predicted score has the same overall mean and standard deviation as the observed composite score.

Data from TOEFL Writing, the *GRE*® General Analytical Writing test (GRE Writing), and the argumentative essay of the *Praxis*® Core Academic Skills for Educators: Writing assessment (Praxis Writing) illustrate use of the proposed methodology in tests that aid in making high-stakes decisions. These testing programs are sufficiently varied to allow examination of the BLP method and its extension via a quadratic penalty function in writing assessments that differ in terms of reported score scale, rate of double human scoring, and number of essay prompts. This report, in great detail, describes the applications of BLP and PBLP methods to these three testing programs, by which means it empirically complements the theoretic framework developed in Haberman et al. (2015) and Yao et al. (2019).

The following two research questions are examined in this study:

1. How well does BLP based on both human holistic scores and computer-generated feature scores perform compared to BLP based solely on human holistic scores in terms of scoring accuracy and assessment accuracy?
2. Can PBLP improve comparability for different subgroups of test scores based on observed composite scores and rescaled test scores based on linear prediction of true composite scores without a major sacrifice of overall scoring accuracy and assessment accuracy?

To apply BLP or PBLP, estimation of variances and covariances of measurement errors are required. As in Haberman et al. (2015) and Yao et al. (2019), we consider two versions of estimation for the variances and covariances of the measurement errors. In the case of scoring accuracy, measurement errors are uncorrelated, and their variances are estimated by use of agreement samples in which more than one rater evaluates an essay. In typical cases of assessment accuracy, the covariance matrix of the measurement error is estimated from repeater data (data on test takers who take a test more than once) weighted by minimum discriminant information adjustment (MDIA), as in Haberman (1984), to compensate for bias due to the usual situation in which repeater samples are not representative of the overall population of test takers.

In the Methods section the basic model for the data is presented, and a brief description is provided of BLP and PBLP. The section Best Linear Predictor and Penalized Best Linear Predictor Methods provides a brief introduction of BLP and PBLP methods. The section Proportional Reduction in Mean Squared Error Measures provides a description of predictor evaluation by mean squared error and proportional reduction in mean squared error (PRMSE), and the section Estimation of Best Linear Predictor and Penalized Best Linear Predictor provides technical details for model estimation. The Scoring Accuracy section treats scoring accuracy, while the Assessment Accuracy section deals with assessment accuracy. The Subgroup Analysis section considers the linear linking procedure applied to subgroup evaluation. The proposed methodology is illustrated in the Applications section through application to three testing programs, and the Discussion section provides concluding remarks.

Methods

This section describes the general measurement model and the most straightforward instances of BLP and PBLP. In the case of BLP, the material described is based on Haberman et al. (2015); however, we consider more general treatment that permits applications to tests with numbers of prompts other than two, whereas the theory of the PBLP method mainly refers to Yao et al.

Consider an assessment with $J \geq 1$ prompts. For some positive integer K, a generic observation is a K-dimensional vector $\mathbf{X}$ with elements X_k, $1 \leq k \leq K$, where $K \geq J$. In the model in which each prompt receives a single human holistic score and no computer-generated data are used, X_k is the holistic score for prompt k, $1 \leq k \leq J = K$. For example, $J = K$ might be 2, where X_1 equals the holistic score assigned by a rater to the first prompt and X_2 equals the holistic score assigned by a rater to the second prompt. Both TOEFL Writing and GRE Writing are special cases of this example; however, in Praxis Writing, $J = K$ reduces to 1. In examples in this report, X_j, $1 \leq j \leq J$, is a human holistic score on prompt j. If $J < K$, then additional human holistic ratings and/or computer-generated features are employed.

In all cases, the vector $\mathbf{X}$ is assumed to have a finite mean and a finite and positive-definite covariance matrix $\text{Cov}(\mathbf{X})$ with row k and column k', $1 \leq k \leq K$ and $1 \leq k' \leq K$, equal to $\text{Cov}\left(X_k, X_{k'}\right)$. The vector $\mathbf{X}$ may be decomposed into an unobserved true score component $\boldsymbol{\tau}$ with elements τ_k, $1 \leq k \leq K$, and an unobserved error component $\mathbf{e}$ with elements e_k, $1 \leq k \leq K$, so that $\mathbf{X} = \boldsymbol{\tau} + \mathbf{e}$. It is assumed that the error vector $\mathbf{e}$ has expectation $\mathbf{0}_K$, where $\mathbf{0}_K$ denotes the K-dimensional variable with all elements 0 and it is assumed that $\mathbf{e}$ has a finite covariance matrix $\text{Cov}(\mathbf{e})$. It is further assumed that the true score $\boldsymbol{\tau}$ and the error $\mathbf{e}$ are uncorrelated, so that τ_k and ${e_k}'$ are uncorrelated for $1 \leq k \leq K$ and $1 \leq k' \leq K$. This assumption implies that the covariance matrix $\text{Cov}(\mathbf{X})$ of the observed vector $\mathbf{X}$ is the sum of the covariance matrix $\text{Cov}(\boldsymbol{\tau})$ of the vector $\boldsymbol{\tau}$ of true scores and the covariance matrix $\text{Cov}(\mathbf{e})$ of the error vector $\mathbf{e}$, so that $\text{Cov}(\mathbf{X}) = \text{Cov}(\boldsymbol{\tau}) + \text{Cov}(\mathbf{e})$.

This report considers estimation of a composite true score $\nu = \mathbf{c}'\boldsymbol{\tau} = \sum_{k=1}^{K} c_k \tau_k$ for some K-dimensional vector $\mathbf{c}$ with elements c_k, $1 \leq k \leq K$. This composite true score corresponds to the composite observed score $O = \mathbf{c}'\mathbf{X}$. Let $E(\mathbf{X})$ denote the expectation of $\mathbf{X}$; then $E(O) = E(\nu) = \mathbf{c}'E(\mathbf{X})$.

TOEFL Writing provides an example with $J = 2$ prompts. One prompt is an integrated task that requires response to a prompt including both an oral and a written stimulus, and another prompt is an independent task that requires a response to a written stimulus. Consider a model in which, for each prompt, there are two human holistic scores and nine computer-generated essay features. In this model, $K = 22$, X_1 equals the first holistic score on the integrated task, X_2 equals the first holistic score on the independent task, X_3 is the second holistic score on the integrated task, X_4 is the second holistic score on the independent prompt, X_5 to X_{13} are the feature variables for the integrated task, and X_{14} to X_{22} are the feature variables for the independent task. If $\mathbf{c}$ is the 22-dimensional vector with elements $c_1 = c_2 = 1/2$ and $c_k = 0$ for $3 \leq k \leq K = 22$, then $\nu = \mathbf{c}'\boldsymbol{\tau}$ is the true score $(\tau_1 + \tau_2)/2$ of the arithmetic mean $(X_1 + X_2)/2$ of the two holistic scores X_1 and X_2. If raters are properly randomized, then $\tau_1 = \tau_3$ and $\tau_2 = \tau_4$, so that $\nu = (\tau_1 + \tau_2)/2$ is also the arithmetic mean of τ_k for $1 \leq k \leq 4$.

Best Linear Predictor and Penalized Best Linear Predictor Methods

In BLP (Haberman et al., 2015; Haberman & Yao, 2015), the best linear predictor $\widehat{\nu} = \alpha + \boldsymbol{\beta}'\mathbf{X}$ for prediction of the composite true score ν by the observed vector $\mathbf{X}$ is the linear function of $\mathbf{X}$ with real intercept α and K-dimensional vector $\boldsymbol{\beta}$ of regression coefficients β_k, $1 \leq k \leq K$ that satisfies $E\left(\left[\nu - \widehat{\nu}\right]^2\right) = \text{MSE}$, where the mean squared error MSE is the minimum of $E([\nu - a - \mathbf{b}'\mathbf{X}]^2)$ for any real a and K-dimensional vector $\mathbf{b}$. As long as the covariance matrix $\text{Cov}(\mathbf{X})$ of $\mathbf{X}$ is positive definite, $\widehat{\nu}$, α, and $\boldsymbol{\beta}$ are uniquely determined and satisfy the equations,

$$\boldsymbol{\beta} = [\text{Cov}(\mathbf{X})]^{-1} \text{Cov}(\mathbf{X}, \nu), \tag{1}$$

and

$$\alpha = E(\nu) - \boldsymbol{\beta}'E(\mathbf{X}), \tag{2}$$

where $\text{Cov}(\mathbf{X}, \nu)$ is the K-dimensional vector with element k, $1 \leq k \leq K$, equal to the covariance of X_k and ν (Rao, 1973, p. 266). The constraints on the covariances of elements of $\boldsymbol{\tau}$ and $\mathbf{e}$ imply that $\text{Cov}(\mathbf{X}, \nu) = \text{Cov}(\boldsymbol{\tau})\mathbf{c}$. This definition of BLP applies no matter what the covariance matrix $\text{Cov}(\mathbf{e})$ of the error vector $\mathbf{e}$ may be. For standardized regression coefficients,

let $\sigma^2(\nu)$ denote the variance $\mathbf{c}'\mathrm{Cov}(\boldsymbol{\tau})\mathbf{c}$ of ν, so that the nonnegative square root $\sigma(\nu)$ of $\sigma^2(\nu)$ is the standard deviation of ν. For $1 \le k \le K$, the standardized regression coefficient corresponding to β_k is $\beta_{*k} = \sigma(X_k)\beta_k/\sigma(\nu)$.

In a more general framework, given a known K_{M} by K design matrix $\mathbf{M}$ of positive rank K_{M}, the predictor $\mathbf{MX}$ is considered. In the simplest case, if $K_{\mathrm{M}} = K$ and $\mathbf{M}$ is the K by K identity matrix, then $\mathbf{MX} = \mathbf{X}$. More generally, if $\mathbf{M}$ is the K_{M} by K matrix with elements $M_{k'k}$ equal to 1 for $k' = k$ and to 0 for $k' \neq k$, then $\mathbf{MX}$ is the vector with elements X_k, $1 \le k \le K_{\mathrm{M}}$. The best linear predictor of ν based on $\mathbf{MX}$ is then $\widehat{\nu}_{\mathrm{M}} = \alpha_{\mathrm{M}} + \boldsymbol{\beta}'_{\mathrm{M}}\mathbf{MX}$, and the residual is $r_{\mathrm{M}} = \nu - \widehat{\nu}_{\mathrm{M}}$, where the K_{M}-dimensional vector $\boldsymbol{\beta}_{\mathrm{M}}$ with elements $\beta_{k\mathrm{M}}$ for $1 \le k \le K_{\mathrm{M}}$ satisfies

$$\boldsymbol{\beta}_{\mathrm{M}} = [\mathrm{Cov}\,(\mathbf{MX})]^{-1}\,\mathbf{M}\mathrm{Cov}\,(\boldsymbol{\tau})\,\mathbf{c} = \left[\mathbf{M}\mathrm{Cov}\,(\mathbf{X})\,\mathbf{M}'\right]^{-1}\mathbf{M}\mathrm{Cov}\,(\boldsymbol{\tau})\,\mathbf{c} \tag{3}$$

and

$$\alpha_{\mathrm{M}} = \left(\mathbf{c} - \mathbf{M}'\boldsymbol{\beta}_{\mathrm{M}}\right)' E\,(\mathbf{X})\,. \tag{4}$$

For $1 \le k \le K_{\mathrm{M}}$, the standardized regression coefficient corresponding to $\beta_{k\mathrm{M}}$ is $\beta_{*k\mathrm{M}} = \sigma(M_k)\beta_{\mathrm{M}k}/\sigma(\nu)$, where M_k is element k of $\mathbf{MX}$.

As previously noted, application of BLP may encounter difficulties when subgroups of examinees must be considered for fairness analysis. For example, gender, race, or ethnicity may be examined in standard testing programs to ensure that specific groups of examinees are not unfairly affected by a particular method of analysis. This issue is a potential concern whenever, for a grouping variable G with positive integer values no greater than some positive integer $H > 1$, knowledge of G affects BLP. Let $G = h$ with positive probability $p_G(h)$ for $1 \le h \le H$. Then concern arises if the conditional MSE $E\left(\left[\nu - \widehat{\nu}\right]^2 | G = h\right)$ given that $G = h$ is not the minimum of the conditional MSE $E([\nu - a - \mathbf{b}'\mathbf{X}]^2 | G = h)$ for real a and K-dimensional vector $\mathbf{b}$ (Haberman & Sinharay, 2013). Assume that the error e_k and the true score ${\tau_k}'$ are sufficiently uncorrelated to the grouping variable G for $1 \le k \le K$ and $1 \le k' \le K$, so that $E(e_k | G = h) = 0$ and $E(e_k {\tau_k}' | G = h) = 0$ for $1 \le h \le H$. Because $E\,(\nu) = E\left(\widehat{\nu}\right)$, the problem with conditional MSE arises if the conditional expectation $E(r | G = h)$ of the residual $r = \nu - \widehat{\nu}$ is not 0 for some positive integer $h \le H$. A measure of the size of this issue is provided by the variance $\sigma^2\left(E\,(r|G)\right) = \sum_{h=1}^{H} p_G\,(h)\left[E\,(r|G = h)\right]^2$ of the random variable $E(r|G)$ with value $E(r|G = h)$ if $G = h$ and $1 \le h \le H$. In PBLP, a constant $d \ge 1$ is selected and $\widehat{\nu}_d = \alpha_d + \boldsymbol{\beta}'_d\mathbf{X}$, α_d real and $\boldsymbol{\beta}_d$ a K-dimensional vector with elements β_{kd} for $1 \le k \le K$, are uniquely defined by the requirement that the residual $r_d = \nu - \widehat{\nu}_d$ satisfies

$$E\left(r_d^2\right) + (d-1)\,\sigma^2\left(E\left(r_d|G\right)\right) = L_d, \tag{5}$$

where L_d is the minimum of $E([\nu - a - \mathbf{b}'\mathbf{X}]^2) + (d-1)E([E(\nu - a - \mathbf{b}'\mathbf{X}|G)]^2)$ for real a and K-dimensional vector $\mathbf{b}$. The case of $d = 1$ is the conventional case of BLP, so that $\widehat{\nu}_1 = \widehat{\nu}$, $\alpha_1 = \alpha$, and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$. For $d > 1$, the penalty $(d-1)\sigma^2(E(r_d|G))$ is assessed to balance the variability of $E(r_d|G)$ against the MSE $\mathrm{MSE}_d = E\left(r_d^2\right)$.

To evaluate $\widehat{\nu}_d$, let $E(\mathbf{X}|G = h)$ be the conditional expectation of $\mathbf{X}$ given $G = h$, and let the conditional expectation $E(\mathbf{X}|G)$ of $\mathbf{X}$ given G be the random vector with value $E(\mathbf{X}|G = h)$ if $G = h$ and $1 \le h \le H$. Let $\mathrm{Cov}(E(\mathbf{X}|\,G)) = \mathrm{Cov}(E(\boldsymbol{\tau}|\,G))$ be the covariance matrix of the conditional expectation $E(\boldsymbol{\tau}|G) = E(\mathbf{X}|G)$ of $\mathbf{X}$ given G. Then

$$\boldsymbol{\beta}_d = [\mathrm{Cov}\,(\mathbf{X}) + (d-1)\,\mathrm{Cov}(E\,(\mathbf{X}|G)]^{-1}\left[\mathrm{Cov}\,(\boldsymbol{\tau}) + (d-1)\,\mathrm{Cov}\,(E\,(\mathbf{X}|G))\right]\mathbf{c} \tag{6}$$

and

$$\alpha_d = \left(\mathbf{c} - \boldsymbol{\beta}_d\right)' E\,(\mathbf{X})\,. \tag{7}$$

The standardized coefficient corresponding to β_{kd} is $\beta_{*kd} = \sigma(X_k)\beta_{kd}/\sigma(\nu)$ for $1 \le k \le K$. If $\mathrm{Cov}(E(\mathbf{X}|\,G))$ is positive definite, then, as d approaches ∞, $\boldsymbol{\beta}_d$ converges to $\mathbf{c}$, α_d converges to 0, $\widehat{\nu}_d$ converges to O, r_d converges to $\nu - O$, and, for $1 \le h \le H$, $E(r_d|G = h)$ converges to 0 and $E\left(\widehat{\nu}_d|G = h\right)$ converges to $E(O|G = h)$. Yao et al. (2019) has provided a detailed derivation of the PBLP method, including Equations 6 and 7.

In the penalty case, similar formulas apply. Thus the predictor $\widehat{\nu}_{\mathrm{M}d} = \alpha_{\mathrm{M}d} + \boldsymbol{\beta}'_{\mathrm{M}d}\mathbf{MX}$ and the residual is $r_{\mathrm{M}d} = \nu - \widehat{\nu}_{\mathrm{M}d}$, where the K_{M}-dimensional vector $\boldsymbol{\beta}_{\mathrm{M}d}$ with elements $\beta_{k\mathrm{M}d}$ for $1 \le k \le K_{\mathrm{K}}$ satisfies

$$\boldsymbol{\beta}_{\mathrm{M}d} = \{\mathbf{M}[\mathrm{Cov}\,(\mathbf{X}) + (d-1)\,\mathrm{Cov}(E\,(\mathbf{X}|G)]\mathbf{M}'\}^{-1}\left\{\mathbf{M}\left[\mathrm{Cov}\,(\boldsymbol{\tau}) + (d-1)\,\mathrm{Cov}\,(E\,(\mathbf{X}|G))\right]\right\}\mathbf{c} \tag{8}$$

and

$$\alpha_{\mathrm{M}d} = \left(\mathbf{c} - \mathbf{M}'\boldsymbol{\beta}_{\mathrm{M}d}\right)' E\,(\mathbf{X})\,. \tag{9}$$

For $1 \leq k \leq K_{\mathbf{M}}$, the standardized coefficient corresponding to $\beta_{k\mathbf{M}d}$ is $\beta_{*k\mathbf{M}d} = \sigma(M_k)\beta_{k\mathbf{M}d}/\sigma(\nu)$. If $\mathbf{c}$ is $\mathbf{M}'\mathbf{c}_{\mathbf{M}}$ for some $K_{\mathbf{M}}$-dimensional vector $\mathbf{c}_{\mathbf{M}}$, $\mathbf{M}\mathrm{Cov}(E(\mathbf{X}|\,G))\mathbf{M}'$ is positive definite, and d approaches ∞, then $\boldsymbol{\beta}_{\mathbf{M}d}$ approaches $\mathbf{c}_{\mathbf{M}}$, $\alpha_{\mathbf{M}d}$ approaches 0, $\widehat{\nu}_{\mathbf{M}d}$ approaches O, $r_{\mathbf{M}d}$ approaches $\nu - O$, and, for $1 \leq h \leq H$, $E\left(\widehat{\nu}_{\mathbf{M}d}|G = h\right)$ approaches $E(O|G = h)$ and $E(r_{\mathbf{M}d}|G = h)$ approaches 0.

In the examples examined in this report, to each prompt j, $1 \leq j \leq J$, correspond 11 potential variables, 2 human holistic scores X_j and X_{J+j}, and 9 essay features $X_{2J+9(j-1)+k}$, $1 \leq k \leq 9$. Thus $K = 11\,J$. If all information is used and the symmetry properties associated with the two holistic scores on each prompt are ignored, then $\mathbf{M}$ is the K by K identity matrix $\mathbf{I}_K$ and $\mathbf{MX} = \mathbf{X}$. If one human score and all feature scores are used for each prompt, then $K_{\mathbf{M}} = K - J$ and $\mathbf{M} = \mathbf{M}_1$, where row k' and column k of $\mathbf{M}_1$ is $M_{k'k1}$ for $1 \leq k' \leq K_{\mathbf{M}}$ and $1 \leq k \leq K$, $M_{k'k1} = 1$ if $k' = k \leq J$ or $k' = k - J$ and $2\,J < k \leq K$, and $M_{k'k1} = 0$ otherwise. If one human score and no computer-generated features are used for each prompt, then $K_{\mathbf{M}} = J$, $\mathbf{M} = \mathbf{M}_2$, row k' and column k of $\mathbf{M}_2$, $1 \leq k' \leq K_{\mathbf{M}}$ and $1 \leq k \leq K$, is $M_{k'k2}$, $M_{k'k2} = 1$ for $k' = k \leq J$ and $M_{k'k2} = 0$ otherwise. If both human scores are used for each prompt, symmetry properties are exploited, and no computer-derived information is used, then $K_{\mathbf{M}} = 2\,J$, $\mathbf{M} = \mathbf{M}_3$, row k' and column k of $\mathbf{M}_3$ is $M_{k'k3}$, $M_{k'k3} = 1/2$ for $k' = k \leq J$ or $k' = k - J$ and $J + 1 \leq k \leq 2\,J$, and $M_{k'k3} = 0$ otherwise. These cases apply to TOEFL Writing and GRE Writing with $J = 2$ and to Praxis Writing with $J = 1$.

Both BLP and PBLP are applied to the two measurement models examined in Haberman et al. (2015). The first definition involves accuracy of scoring. In the example with four human scores and 18 computer-derived essay features, electronic scoring involves no errors in the sense that the computer program always gives the same feature values to the same essay. As a consequence, the variances $\sigma^2(e_k)$ of the errors e_k are 0 for $5 \leq k \leq K = 20$. In addition, all errors e_k, $1 \leq k \leq 4$, are uncorrelated, $\sigma^2(e_1) = \sigma^2(e_3)$, and $\sigma^2(e_2) = \sigma^2(e_4)$. The other case involves accuracy of assessment, where the prompts are regarded as drawn from pools of comparable prompts, so that the weighted average true score $c_1\tau_1 + c_2\tau_2$ can be regarded as the expected weighted average of human holistic scores among parallel tests. In this case, the variances $\sigma^2(e_k)$ can be assumed to be positive for $1 \leq k \leq K$, for computers can be expected to give different feature scores to different essays. The errors of essay feature variables may be correlated with one another; that is, the covariances $\mathrm{Cov}\left(e_k, e_{k'}\right)$ may be nonzero if $1 \leq k < k' \leq 20$. The assumption (Haberman et al., 2015) is not needed that errors for different prompts are uncorrelated. It is also assumed that $\mathrm{Cov}(e_1, e_k) = \mathrm{Cov}(e_3, e_k)$ and $\mathrm{Cov}(e_2, e_k) = \mathrm{Cov}(e_4, e_k)$ for $5 \leq k \leq 20$, $\sigma^2(e_1) = \sigma^2(e_3)$, $\sigma^2(e_2) = \sigma^2(e_4)$, and $\mathrm{Cov}(e_1, e_2) = \mathrm{Cov}(e_3, e_2) = \mathrm{Cov}(e_1, e_4) = \mathrm{Cov}(e_3, e_4)$. This model applies to TOEFL Writing and GRE Writing.

If, as in Praxis Writing, only one prompt is considered in the writing assessment, $K = 11$, X_1 equals the first holistic score on the prompt, X_2 is the second holistic score on the prompt, and X_3 to X_{11} are the feature variables for this prompt. If $\mathbf{c}$ is the 11-dimensional vector with elements $c_1 = 1$ and $c_k = 0$ for $2 \leq k \leq 11$, then $\nu = \mathbf{c}'\boldsymbol{\tau}$ is the true score τ_1 of the first holistic score. Given proper rater randomization, $\tau_1 = \tau_2$. In the case of accuracy of scoring, the variances $\sigma^2(e_k)$ of the errors e_k are 0 for $3 \leq k \leq K = 11$, the errors e_1 and e_2 are uncorrelated, and $\sigma^2(e_1) = \sigma^2(e_2)$. Although available data do not permit evaluation of accuracy of assessment, were such data attainable, the variances $\sigma^2(e_k)$ can be assumed to be positive for $1 \leq k \leq K$, and the covariances $\mathrm{Cov}\left(e_k, e_{k'}\right)$ may be nonzero if $1 \leq k < k' \leq 11$. For human scores and essay feature scores, $\mathrm{Cov}(e_1, e_k) = \mathrm{Cov}(e_2, e_k)$ for $3 \leq k \leq 11$ and $\sigma^2(e_1) = \sigma^2(e_2)$.

Proportional Reduction in Mean Squared Error Measures

To evaluate the quality of various sets of predictors, the PRMSE $\rho^2 = 1 - \mathrm{MSE}/\sigma^2(\nu)$ provides a measure of the effectiveness of best linear prediction by the complete vector $\mathbf{X}$. In other words, ρ^2 is the coefficient of determination from prediction of ν by the linear predictor $\widehat{\nu}$ relative to prediction of ν by the constant $E(\nu)$ (Haberman, 2008). The coefficient ρ^2 is nonnegative and does not exceed 1. Larger values of ρ^2 indicate more effective prediction. In the Applications section, ρ^2 is applied to evaluate three sets of predictors. In examination of ρ^2, the following equations are often helpful:

$$\mathrm{MSE} = \mathbf{c}'\left\{\mathrm{Cov}\left(\boldsymbol{\tau}\right) - \mathrm{Cov}\left(\boldsymbol{\tau}\right)\left[\mathrm{Cov}\left(\mathbf{X}\right)\right]^{-1}\mathrm{Cov}\left(\boldsymbol{\tau}\right)\right\}\mathbf{c}, \tag{10}$$

and

$$\rho^2 = \frac{\mathbf{c}'\mathrm{Cov}\left(\boldsymbol{\tau}\right)\left[\mathrm{Cov}\left(\mathbf{X}\right)\right]^{-1}\mathrm{Cov}\left(\boldsymbol{\tau}\right)\mathbf{c}}{\mathbf{c}'\mathrm{Cov}\left(\boldsymbol{\tau}\right)\mathbf{c}}. \tag{11}$$

In the more general framework, for the linear predictor $\widehat{\nu}_{\mathbf{M}}$, the mean squared error $\mathrm{MSE}_{\mathbf{M}} = E\left(r_{\mathbf{M}}^2\right)$, and the coefficient of determination $\rho_{\mathbf{M}}^2 = 1 - \mathrm{MSE}_{\mathbf{M}}/\sigma^2(\nu)$. Here

$$\mathrm{MSE}_{\mathbf{M}} = \mathbf{c}'\left\{\mathrm{Cov}(\boldsymbol{\tau}) - \mathrm{Cov}(\boldsymbol{\tau})\,\mathbf{M}'\left[\mathrm{Cov}(\mathbf{MX})\right]^{-1}\mathbf{M}\mathrm{Cov}(\boldsymbol{\tau})\right\}\mathbf{c} \tag{12}$$

and

$$\rho_{\mathbf{M}}^2 = \frac{\mathbf{c}'\mathrm{Cov}(\boldsymbol{\tau})\,\mathbf{M}'\left[\mathrm{Cov}(\mathbf{MX})\right]^{-1}\mathbf{M}\mathrm{Cov}(\boldsymbol{\tau})\,\mathbf{c}}{\mathbf{c}'\mathrm{Cov}(\boldsymbol{\tau})\,\mathbf{c}}. \tag{13}$$

For $\mathbf{M}$ equal to the K by K identity matrix, $\alpha_{\mathbf{M}} = \alpha$, $\boldsymbol{\beta}_{\mathbf{M}} = \boldsymbol{\beta}$, $\mathrm{MSE}_{\mathbf{M}} = \mathrm{MSE}$, and $\rho_{\mathbf{M}}^2 = \rho^2$. In general, $\rho_{\mathbf{M}}^2 \le \rho^2$.

When a penalty function is employed and $d > 1$, then MSE increases in typical cases. For the complete predictor $\mathbf{X}$, $\mathrm{MSE}_d = E\left(r_d^2\right) \ge \mathrm{MSE}$ and $\sigma^2(E(r_d|G)) \le \sigma^2(E(r|G))$, with equality only if the conditional expectation $E(r|G)$ of the residual r from BLP is always 0. Similarly, $\mathrm{MSE}_{\mathbf{M}d} = E\left(r_{\mathbf{M}d}^2\right) \ge \mathrm{MSE}_{\mathbf{M}}$ and $\sigma^2(E(r_{\mathbf{M}d})) \le \sigma^2(E(r_{\mathbf{M}}|G))$, with equality only if $E(r_{\mathbf{M}}|G)$ is always 0. In general, $\mathrm{MSE}_{\mathbf{M}d} \ge \mathrm{MSE}_d$, with equality only if $\widehat{\nu}_d = \widehat{\nu}_{\mathbf{M}d}$. The proportional reductions in MSE are $\rho_d^2 = 1 - \mathrm{MSE}_d/\sigma^2(\nu)$ and $\rho_{\mathbf{M}d}^2 = 1 - \mathrm{MSE}_{\mathbf{M}d}/\sigma^2(\nu)$. It must be the case that $0 \le \rho_{\mathbf{M}d}^2 \le \rho_d^2 \le 1$.

Estimation of Best Linear Predictor and Penalized Best Linear Predictor

To estimate α, $\alpha_{\mathbf{M}}$, $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{\mathbf{M}}$, MSE, $\mathrm{MSE}_{\mathbf{M}}$, ρ^2, and $\rho_{\mathbf{M}}^2$, the expectation $E(\mathbf{X}) = E(\boldsymbol{\tau})$ and the covariance matrices $\mathrm{Cov}(\mathbf{X})$ and $\mathrm{Cov}(\boldsymbol{\tau})$ must be estimated. In addition, estimation of the corresponding parameters for a penalty function with $d > 1$ requires estimation of the conditional expectation $E(\mathbf{X}|G = h)$ of $\mathbf{X}$ given $G = h$ and the conditional covariance matrix $\mathrm{Cov}(E(\mathbf{X}|G))$. To ensure that estimated covariance matrices can be nonsingular, assume that the sample size n exceeds the dimension K of $\mathbf{X}$. To accommodate test takers who take the assessment more than once (repeaters), a slightly more complex sampling procedure is required than is usually considered. Let the observations correspond to T test takers, where T may be less than n. For test taker t from 1 to T, let $N(t)$ be the set of observations corresponding to that test taker, and, for each observation i, $1 \le i \le n$, let $t(i)$ denote the corresponding test taker. Assume that the testing program has a maximum number of times a test taker can take the assessment during the time period represented by the sample, so that the maximum number of members of $N(t)$ is bounded. For $1 \le t \le T$, let $R(t)$ be the number of elements of $N(t)$, and let $i_*(i', t)$, $1 \le i' \le R(t)$, be the observation that corresponds to the i'th test administration taken by test taker t. It is assumed that T, $R(t)$, $1 \le t \le T$, and $i_*(i', t)$, $1 \le i' \le R(t)$ and $1 \le t \le T$, are all random variables. For observation i, let $\mathbf{X}_i$ be a K-dimensional random vector with elements X_{ik}, $1 \le k \le K$, such that each $\mathbf{X}_i$ has the same distribution as $\mathbf{X}$. For $1 \le t \le T$, let $\tilde{\mathbf{X}}_t$ be the K by $R(t)$ matrix with column i', $1 \le i' \le R(t)$, equal to $\mathbf{X}_{i_*(i',t)}$. Let the $\tilde{\mathbf{X}}_t$, $1 \le t \le T$, be independently distributed, and let the conditional distribution of $\tilde{\mathbf{X}}_t$ and $\tilde{\mathbf{X}}_{t'}$ given $R(t) = R(t')$ be the same if $1 \le t < t' \le T$.

The estimate of the expectation $E(\mathbf{X})$ is the sample mean

$$\overline{E}(\mathbf{X}) = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i, \tag{14}$$

and the covariance matrix $\mathrm{Cov}(\mathbf{X})$ is estimated by the (biased) sample covariance matrix

$$\overline{\mathrm{Cov}}(\mathbf{X}) = n^{-1}\sum_{i=1}^{n}\left[\mathbf{X}_i - \overline{E}(\mathbf{X})\right]\left[\mathbf{X}_i - \overline{E}(\mathbf{X})\right]'. \tag{15}$$

The vector $\overline{E}(\mathbf{X})$ has elements

$$\overline{E}\left(X_k\right) = n^{-1}\sum_{i=1}^{n}X_{ik} \tag{16}$$

for $1 \le k \le K$, and $\overline{\mathrm{Cov}}(\mathbf{X})$ has row k and column k' equal to

$$\overline{\mathrm{Cov}}\left(X_k, X_{k'}\right) = n^{-1}\sum_{i=1}^{n}\left[X_{ik} - \overline{E}\left(X_k\right)\right]\left[X_{ik'} - \overline{E}\left(X_{k'}\right)\right] \tag{17}$$

for positive integers k and k' not greater than K.

In applications in this report, some modification of these formulas is needed due to failure to observe $X_{i(J+j)}$, $1 \le j \le J$, for most observations i, $1 \le i \le n$. The details of the modification are described in Appendix A.

Modifications are mostly important in this report for $\mathbf{M} = \mathbf{I}_K$ and $\mathbf{M} = \mathbf{M}_3$. They help explore expected changes in scoring performance if double human scoring is employed.

In the case of PBLP with $d > 1$, further estimation is needed. Let examinee i be in group G_i, and let $\overline{E}(\mathbf{X}|G = h)$ be the average of $\mathbf{X}_i$ for examinees i with $G_i = h$, $1 \le h \le H$. It is assumed that $\overline{p}_G(h)$, the fraction of observations $i \le n$ with $G_i = h$, is positive for $1 \le h \le H$. Then $\mathrm{Cov}(E(\mathbf{X}|\,G))$ has estimate

$$\begin{aligned}\overline{\mathrm{Cov}}\left(E\left(\mathbf{X}|G\right)\right) &= n^{-1}\sum_{i=1}^{n}\left[\overline{E}\left(\mathbf{X}|G = G_i\right) - \overline{E}\left(\mathbf{X}\right)\right]\left[\overline{E}\left(\mathbf{X}|G = G_i\right) - \overline{E}\left(\mathbf{X}\right)\right]' \\ &= \sum_{h=1}^{H}\overline{p}_G(h)\left[\overline{E}\left(\mathbf{X}|G = h\right) - \overline{E}\left(\mathbf{X}\right)\right]\left[\overline{E}\left(\mathbf{X}|G = h\right) - \overline{E}\left(\mathbf{X}\right)\right]'. \end{aligned} \tag{18}$$

Replacement of $\overline{E}(\mathbf{X}|G = h)$ by $\overline{E}\left(\widetilde{\mathbf{X}}|G = h\right)$ for $1 \le h \le H$ and $\overline{\mathrm{Cov}}\left(E\left(\mathbf{X}|G\right)\right)$ by $\overline{\mathrm{Cov}}\left(E\left(\widetilde{\mathbf{X}}|G\right)\right)$ is appropriate in applications in this report in which $X_{i(J+j)}$ is only available for a subset i in U_j for $1 \le j \le J$.

It is far more difficult to estimate the covariance matrix $\mathrm{Cov}(\boldsymbol{\tau})$ of the vector $\boldsymbol{\tau}$ of true scores. Two types of estimates are considered in this report: scoring accuracy and assessment accuracy. In each of these cases, an estimate $\overline{\mathrm{Cov}}(\boldsymbol{\tau})$ of $\mathrm{Cov}(\boldsymbol{\tau})$ is provided, so that all desired parameters may be estimated by use of obvious substitutions such as $\overline{E}(\mathbf{X})$ or $\overline{E}\left(\widetilde{\mathbf{X}}\right)$ for $E(\mathbf{X})$, $\overline{\mathrm{Cov}}(\mathbf{X})$ or $\overline{\mathrm{Cov}}_m(\mathbf{X})$ for $\mathrm{Cov}(\mathbf{X})$, $\overline{\mathrm{Cov}}(\boldsymbol{\tau})$ for $\mathrm{Cov}(\boldsymbol{\tau})$, $\overline{E}(\mathbf{X}|G = h)$ or $\overline{E}\left(\widetilde{\mathbf{X}}|G = h\right)$ for $E(\mathbf{X}|G = h)$, $1 \le h \le H$, and $\overline{\mathrm{Cov}}\left(E\left(\mathbf{X}|G\right)\right)$ or $\overline{\mathrm{Cov}}\left(E\left(\widetilde{\mathbf{X}}|G\right)\right)$ for $\mathrm{Cov}(E(\mathbf{X}|\,G))$.

Scoring Accuracy

In the case of scoring accuracy when only agreement samples are available for $X_{i(J+j)}$, $1 \le j \le J$, the variance of measurement error $\sigma^2(e_j) = \sigma^2(e_{J+k})$ has unbiased estimate $\overline{\sigma}^2\left(e_j\right) = \overline{\sigma}^2\left(e_{J+j}\right) = \overline{\sigma}^2\left(X_j - X_{J+j}\right)/2$. Because $\mathrm{Cov}\left(e_k, e_{k'}\right)$ is 0 for $k \ne k'$ and for k or k' greater than $2J$, estimation of $\mathrm{Cov}(\mathbf{e})$ is straightforward. Let $\overline{\mathrm{Cov}}(\mathbf{e})$ be the diagonal matrix with kth diagonal element $\overline{\sigma}^2\left(e_j\right)$ for $1 \le j \le 2J$ and other diagonal elements 0. Then $\mathrm{Cov}(\boldsymbol{\tau})$ has estimate

$$\overline{\mathrm{Cov}}\left(\boldsymbol{\tau}\right) = \overline{\mathrm{Cov}}_m\left(\mathbf{X}\right) - \overline{\mathrm{Cov}}\left(\mathbf{e}\right). \tag{19}$$

Assessment Accuracy

Study of assessment accuracy typically involves use of repeater data for estimation of the covariance matrix of the vector of true scores (Haberman et al., 2015). In addition to the K-dimensional random vector $\mathbf{X}$, an additional K-dimensional random vector $\mathbf{X}_*$ with elements X_{ik^*}, $1 \le k \le K$, represents test results that the same examinee would have had if that test taker had taken a different parallel assessment instead of the one taken. Under this ideal condition, $(\mathbf{X}, \mathbf{X}_*)$ and $(\mathbf{X}_*, \mathbf{X})$ have the same distribution. It follows that $\mathrm{Cov}(\boldsymbol{\tau})$ is the K by K matrix with row k and column k' equal to $\mathrm{Cov}\left(X_k, X_{k'*}\right)$ for positive integers k and k' not greater than K. Under typical conditions, this approach to repeater data is difficult to apply because repeater data are normally only available on a nonrepresentative subset of the complete sample and because repeating a test may well affect performance. To treat this problem of unrepresentative sampling, MDIA may be used (Haberman, 1984). In this approach, sample weights are employed so that the weighted sample of repeaters satisfies a finite set of constraints on weighted averages. The sample MDIA weights minimize the sample discriminant information subject to the given constraints. The discussion here follows Haberman (1984), Haberman et al. (2015), and Haberman and Yao (2015), and the procedure is described in Appendix B.

Subgroup Analysis

The basic tool for examination of behavior of a linear predictor when applied to a subgroup is the conditional expectation of the residual. For the general prediction vector $\mathbf{X}$, the conditional expectation $E(r|G = h)$ of the residual r given $G = h$

has estimate $\overline{E}(r|G=h)=\overline{E}(O|G=h)-\widehat{\boldsymbol{\beta}}'\overline{E}(\mathbf{X}|G=h)$ for $1\le h\le G$, where $\widehat{\boldsymbol{\beta}}$ denotes the estimated value of $\boldsymbol{\beta}$. Similarly, for $d\ge 1$, $E(r_d|G=h)$ has estimate $\overline{E}\left(r_d|G=h\right)=\overline{E}(O|G=h)-\widehat{\boldsymbol{\beta}}_d'\overline{E}(\mathbf{X}|G=h)$, where $\widehat{\boldsymbol{\beta}}_d$ is the estimate of $\boldsymbol{\beta}_d$. For the matrix $\mathbf{M}$, $E(r_{\mathbf{M}}|G=h)$ has estimate $\overline{E}\left(r_{\mathbf{M}}|G=h\right)=\overline{E}(O|G=h)-\widehat{\boldsymbol{\beta}}_{\mathbf{M}}'\overline{E}(\mathbf{X}|G=h)$, where $\widehat{\boldsymbol{\beta}}_{\mathbf{M}}$ estimates $\boldsymbol{\beta}_{\mathbf{M}}$. For the matrix $\mathbf{M}$ and $d\ge 1$, $\overline{E}\left(r_{\mathbf{M}d}|G=h\right)=\overline{E}(O|G=h)-\widehat{\boldsymbol{\beta}}_{\mathbf{M}d}'\overline{E}(\mathbf{X}|G=h)$ estimates $E(r_{\mathbf{M}d}|G=h)$, where $\widehat{\boldsymbol{\beta}}_{\mathbf{M}d}$ estimates $\boldsymbol{\beta}_{\mathbf{M}d}$. In all cases, it is desirable that the conditional expectation of residuals be small. In the case of agreement samples, $\widetilde{\mathbf{X}}$ replaces $\mathbf{X}$.

A complication in the analysis involves the definition of what is a small value. An added complication is that in normal testing practice, a predictor $\widehat{\nu}$, $\widehat{\nu}_d$, $\widehat{\nu}_{\mathbf{M}}$, or $\widehat{\nu}_{\mathbf{M}d}$ is converted to a scale score before a test result is reported. Conversion procedures vary considerably even for tests within the same company, but one possibility to examine this issue involves use of a standardized scaling associated with linear equating. For $d\ge 1$ and a matrix $\mathbf{M}$, consider the following scaled versions of $\widehat{\nu}_d$ and $\widehat{\nu}_{\mathbf{M}d}$:

$$\widehat{\zeta}_d=E(O)+\frac{\sigma(O)}{\sigma\left(\widehat{\nu}_d\right)}\left[\widehat{\nu}_d-E(O)\right] \tag{20}$$

and

$$\widehat{\zeta}_{\mathbf{M}d}=E(O)+\frac{\sigma(O)}{\sigma\left(\widehat{\nu}_{\mathbf{M}d}\right)}\left[\widehat{\nu}_{\mathbf{M}d}-E(O)\right]. \tag{21}$$

Here the observed composite O has standard deviation $\sigma(O)=[\mathbf{c}'\mathrm{Cov}(\mathbf{X})\mathbf{c}]^{1/2}$, $\widehat{\nu}_d$ has standard deviation $\sigma\left(\widehat{\nu}_d\right)=\left[\boldsymbol{\beta}_d'\mathrm{Cov}(\mathbf{X})\boldsymbol{\beta}_d\right]^{1/2}$, and $\widehat{\nu}_{\mathbf{M}d}$ has standard deviation $\left[\boldsymbol{\beta}_{\mathbf{M}d}'\mathbf{M}\mathrm{Cov}(\mathbf{X})\mathbf{M}'\boldsymbol{\beta}_{\mathbf{M}d}\right]^{1/2}$. The scaled values $\widehat{\zeta}_d$ and $\widehat{\zeta}_{\mathbf{M}d}$ have the same mean $E(O)$ and standard deviation $\sigma(O)$ as the observed composite score O. In ideal situations, $E\left(\widehat{\zeta}_d|G=h\right)=E(O|G=h)$ for $1\le h\le H$. This result always holds for $\widehat{\zeta}_{\mathbf{M}_0d}$ if $K_{\mathbf{M}}=1$, $M_{1k0}=1$ for $1\le k\le J$, and $M_{1k0}=0$ for $k>J$, for the linear predictor $\widehat{\nu}_{\mathbf{M}_0d}$ is a linear function of the observed composite score O. In general, each conditional expectation $E\left(\widehat{\zeta}_d|G=h\right)=E(O|G=h)$ if each expected conditional residual $E(r_d|G=h)$, $1\le h\le H$, is $\left[1-\sigma\left(\widehat{\nu}_d\right)/\sigma(O)\right]\left[E(O|G=h)-E(O)\right]$. If $\mathrm{Cov}(E(\mathbf{X}|G))$ is positive definite and d approaches ∞, then $E\left(\widehat{\zeta}_d|G=h\right)$ converges to $E(O|G=h)$ for $1\le h\le H$. Each conditional expectation $E\left(\widehat{\zeta}_{\mathbf{M}d}|G=h\right)$ is $E(O|G=h)$ if each expected conditional residual $E(r_{\mathbf{M}d}|G=h)$ is $[1-\sigma(\widehat{\nu}_{\mathbf{M}d}/\sigma(O)]\left[E(O|G=h)-E(O)\right]$. If $\mathbf{M}\mathrm{Cov}(\mathbf{X}|G))\mathbf{M}'$ is positive definite, $\mathbf{c}=\mathbf{M}'\mathbf{c}_{\mathbf{M}}$ for a $K_{\mathbf{M}}$-dimensional vector $\mathbf{c}_{\mathbf{M}}$, and d approaches ∞, then $E\left(\widehat{\zeta}_{\mathbf{M}d}|G=h\right)$ converges to $E(O|G=h)$ for $1\le h\le H$. The convention is adopted that $\widehat{\zeta}=\widehat{\zeta}_1$ corresponds to the original BLP $\widehat{\nu}$, and $\widehat{\zeta}_{\mathbf{M}}=\widehat{\zeta}_{\mathbf{M}1}$ corresponds to $\widehat{\nu}_{\mathbf{M}}$.

Estimates of $E\left(\widehat{\zeta}_d|G=h\right)$ and $E\left(\widehat{\zeta}_{\mathbf{M}d}|G=h\right)$ are obtained by substitution of $\overline{E}(\mathbf{X}|G=h)$ or $\overline{E}\left(\widetilde{\mathbf{X}}|G=h\right)$ for $E(\mathbf{X}|G=h)$, $\overline{\mathrm{Cov}}(\mathbf{X})$ or $\overline{\mathrm{Cov}}_m\left(\widetilde{\mathbf{X}}\right)$ for $\mathrm{Cov}(\mathbf{X})$, $\overline{\mathrm{Cov}}(\boldsymbol{\tau})$ for $\mathrm{Cov}(\boldsymbol{\tau})$, and $\overline{\mathrm{Cov}}(E(\mathbf{X}|G))$ or $\overline{\mathrm{Cov}}\left(E\left(\widetilde{\mathbf{X}}|G\right)\right)$ for $\mathrm{Cov}(E(\mathbf{X}|G))$.

Applications

Data Summary

This study involves data collected from three large-scale assessments used to make high-stakes decisions, namely, TOEFL Writing, GRE Writing, and Praxis Writing, administered by Educational Testing Service. Both TOEFL iBT and GRE writing tests include two essay prompts, and the portion of the Praxis writing assessment under consideration in this study contains only one essay prompt. For TOEFL Writing, the independent prompt asks the test takers to express their opinions on a subject, and the integrated prompt asks the test takers to integrate information from a reading source and a listening source. In GRE Writing, test takers are asked to write an argumentative essay on a topic with reasons and supporting evidence (issue) and an essay to evaluate an argument (argument). Despite the name, the argumentative prompt in Praxis Writing is similar to the issue prompt in GRE Writing. The data set for each assessment was collected in different periods of time. Specifically, the TOEFL main sample consists of all essay responses collected from 1,006,554 examinees who took TOEFL iBT tests between January 9 and December 20, 2015; the GRE main sample consists of responses randomly selected from 194,851 test-takers' essay responses between July 1, 2013, and June 30, 2016; and the Praxis main sample consists of all responses of 149,713 examinees between October 6, 2014, and December 31, 2016.

For all programs, each essay response was graded by a randomly chosen but trained human rater. For each prompt, there is a predetermined percentage of essays that were randomly selected and scored by a second operational rater. This sample is the agreement sample. Although the rate of double rating varied by program, in all cases, the agreement sample was just a small portion of the full sample for each prompt. It is also noted that the sample sizes vary not only by different testing programs but also by prompts. For TOEFL, the agreement sample contains 44,381 responses for the independent prompt and 44,831 responses for the integrated prompt. For GRE, the agreement sample includes 8,988 responses for the issue prompt and 8,306 for the argument prompt. The sample size of the agreement sample for Praxis Writing is 10,990. The data of repeaters who took the same test more than once were also selected in each prompt, and obviously the repeater data were just a subsample of the main population: 168,595 for TOEFL, 4,739 for GRE, and 23,673 for Praxis, respectively. Lastly, we also have subgroup information available for each data set under study, that is, 19 subgroups defined for TOEFL, 17 for GRE, and 7 for Praxis. It is worthwhile noting that the subgroup compositions are characterized in different ways by testing program. For TOEFL Writing, the subgroups are defined in terms of a combination of test region (geographic location of where the test takes place) and native language information. For GRE Writing, the subgroups are defined using test region for test takers outside the United States, and test takers in the United States are further divided by ethnicity information into seven subgroups (e.g., White, Asian, Black, Hispanic). Finally, the subgroups are divided based on ethnicity information for Praxis Writing.

In addition to human scores, we also obtained e-rater feature scores extracted from each essay. The e-rater engine uses natural language processing techniques to extract information of essays to compute feature scores (Burstein et al., 2004). In this study, we used nine e-rater feature scores in the prediction models (Attali et al., 2003; Attali & Burstein, 2006; Burstein et al., 2004; Haberman & Sinharay, 2010) as follows:

1. *grammar*: minus the square root of the number of grammatical errors detected per word
2. *usage*: minus the square root of the number of usage errors detected per word
3. *mechanics*: minus the square root of the number of mechanics errors detected per word
4. *vocabulary sophistication*: minus the median Standard Frequency Index value of words not excluded by search engines
5. *word complexity*: the average number of characters per word
6. *syntactic variety*: a measure of the diversity of syntactic structure of the sentences in an essay
7. *development*: the logarithm of the average number of words per discourse element
8. *organization*: the logarithm of the number of discourse elements
9. *collocation-preposition*: a measure of correctness of use of collocations and prepositions encountered in everyday English vocabulary

Results of Applications

In all cases, $c_k = 0$ for $k > J$ and X_j, $1 \le j \le J$, was a human holistic score for prompt j. Three linear predictors of ν were obtained with three selections of $\mathbf{MX}$. Composite true score $\sum_{k=1}^{J} c_k \tau_k$ was predicted by the BLP based on $\mathbf{MX}$. The following matrices were examined:

- Matrix 1 ($\mathbf{M}_1$): For each prompt, one human rating and nine e-rater features were used.
- Matrix 2 ($\mathbf{M}_2$): For each prompt, one human rating was used.
- Matrix 3 ($\mathbf{M}_3$): For each prompt, the average of two human ratings was used.

Two choices of c_k, $1 \le k \le J$, were considered when $J > 1$, as is the case for TOEFL Writing and GRE Writing. The first choice used equal weighting, so that each $c_k = J^{-1}$. In the second choice, c_k is proportional to the inverse of the standard deviation of the true score τ_k of X_k. If $J = 1$, as is the case for Praxis Writing, then c_1 must be 1.

Results are reported for scoring accuracy in the section Results for Scoring Accuracy (Research Question 1) and for assessment accuracy in the section Results for Assessment Accuracy (Research Question 1). Yao et al. (2019) presented the results of PRMSE measures for different models; however, there are more substantial results to validate the psychometric quality of computer-generated feature scores in the proposed BLP models that are worth disseminating. For this purpose, the explicit explanation illustrated in these two sections well answers the first research question. The effectiveness of the penalty function based on different strength of parameter d is examined in the section Subgroup Analyses Based on

Penalized Best Linear Predictor Method (Research Question 2). Yao et al. (2019) gives the results of subgroup analyses in the case of scoring accuracy for the three writing assessments. To provide a more complete picture, the section Subgroup Analyses Based on Penalized Best Linear Predictor Method (Research Question 2) complements the analyses in Yao et al. (2019) by adding the results for assessment accuracy as well as a thorough comparison between these two sets of results.

Results for Scoring Accuracy (Research Question 1)

In the study of scoring accuracy, the agreement samples are used for estimating the variance of the true scores and the measurement errors. The estimates of the PRMSE ρ_{M^2} for all three tests and for both choices of weights (when $J > 1$, i.e., TOEFL and GRE) and only one weight (when $J = 1$, i.e., Praxis) indicate how accurate the tests measure the attribute of interest.

With the three matrices $\mathbf{M}$, for TOEFL Writing and GRE Writing, there are relatively small differences in estimated PRMSEs for the two weighting selections. This situation is especially true for GRE Writing, where the two weights c_1 and c_2 are quite close (.5152 and .4848 for issue and argument prompts, respectively). The weights are .6126 and .3874 for independent and integrated prompts, respectively, in TOEFL Writing. In general, with matrix $\mathbf{M}_2$ (i.e., only the first human score for each prompt as the predictor), the estimated PRMSE statistics are the lowest, that is, the values are .8398 with equal weighting and .8271 with unequal weighting for TOEFL, .8528 with equal weighting and .8528 with unequal weighting for GRE, and .7139 for Praxis. A comparison of $\mathbf{M}_2$ to $\mathbf{M}_1$ shows that using e-rater features provides an appreciable increase in estimated PRMSE. With matrix $\mathbf{M}_1$ (i.e., the first human score and nine e-rater feature scores for each prompt as the predictors), the estimated PRMSE statistics rise to .8864 with equal weighting and .8856 with unequal weighting for TOEFL, .9184 with equal weighting and .9189 with unequal weighting for GRE, and .9300 for Praxis. A comparison of $\mathbf{M}_3$ to $\mathbf{M}_2$ indicates the effectiveness of the employment of the second human rating. Note that, with matrix $\mathbf{M}_3$ (i.e., the average of two human scores for each prompt as the predictor), the estimated PRMSE statistics further go up to .9119 with equal weighting and to .9022 with unequal weighting for TOEFL, and to .9205 with equal weighting and to .9205 with unequal weighting for GRE, but only increase to .8331 for Praxis. In the case of TOEFL Writing, $\mathbf{M}_1$ performs less well than $\mathbf{M}_3$, whereas in GRE Writing, $\mathbf{M}_1$ yields results only slightly worse than for $\mathbf{M}_3$. However, the results for Praxis Writing are somewhat different from the ones for TOEFL Writing and GRE Writing, for $\mathbf{M}_1$ clearly outperforms the other two matrices in terms of estimated PRMSE. In this case, the employment of machine feature scores yields a substantially higher PRMSE estimate compared to exclusive use of human scores ($\mathbf{M}_2$ and $\mathbf{M}_3$).

Estimated standardized partial regression coefficients can shed light on the relative importance of the predictors in the proposed prediction models. As an illustration, consider TOEFL Writing. Table 1 shows the estimated standardized regression coefficients for three matrices and two choices of weights. Results for equal weights show that, with $\mathbf{M}_1$, the human ratings from the integrated prompt receive by far the highest weight of any of the predictors, followed by the human ratings from the independent prompt.

Among the e-rater features, the development and organization features, especially the ones for the independent prompt, receive much larger weights than the other features. Using additional human ratings ($\mathbf{M}_3$) does not appear to alter the weights for the human ratings in either essay prompt, as is evident from the comparable weights between $\mathbf{M}_2$ and $\mathbf{M}_3$. However, the weights for the human ratings on both prompts are notably decreased when e-rater features are added as predictors. In particular, the weight for human ratings on the integrated prompt dropped nearly by half from $\mathbf{M}_2$ and $\mathbf{M}_3$ to $\mathbf{M}_1$. With unequal weights, the general pattern of the relative weights for each predictor with different matrices is similar to the results with equal weights. Nonetheless, because the standard deviations of τ_k for the two prompts are quite different, when weights are unequal, estimated standardized regression coefficients for human ratings on the integrated prompt are somewhat smaller and corresponding coefficients are somewhat larger on the independent prompt.

Comparable results for GRE Writing are shown in Table 2. For both weighting procedures, the human ratings from the argument prompt receive the highest estimated standardized regression coefficients with $\mathbf{M}_1$. Interestingly, for both tests, the human ratings on essay prompts that evaluate more complex, higher order skills (e.g., argumentation, reading and listening), such as the integrated prompt in TOEFL Writing and the argument prompt in GRE Writing, have the largest estimated standardized regression coefficients. In addition, similar to the results in TOEFL Writing, the estimated standardized regression coefficient for the human ratings on the other prompt (issue) decreases substantially to nearly half of the original value when e-rater features are added.

Table 1 Estimated Standardized Regression Coefficients for Scoring Accuracy: TOEFL Writing

Weight	Variable	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_3$
Equal	Human rating: Independent	.2058	.3580	.3753
	Human rating: Integrated	.5521	.6814	.6846
	Grammar: Independent	.0678	–	–
	Usage: Independent	.0487	–	–
	Mechanics: Independent	.0438	–	–
	Vocabulary sophistication: Independent	.0421	–	–
	Word complexity: Independent	.0171	–	–
	Syntactic variety: Independent	.0283	–	–
	Development: Independent	.1110	–	–
	Organization: Independent	.1263	–	–
	Collocation-preposition: Independent	.0155	–	–
	Grammar: Integrated	.0390	–	–
	Usage: Integrated	.0136	–	–
	Mechanics: Integrated	.0210	–	–
	Vocabulary sophistication: Integrated	.0305	–	–
	Word complexity: Independent	−.0065	–	–
	Syntactic variety: Integrated	.0314	–	–
	Development: Integrated	.1065	–	–
	Organization: Integrated	.0961	–	–
	Collocation-preposition: Integrated	.0085	–	–
Unequal	Human rating: Independent	.2469	.4204	.4563
	Human rating: Integrated	.4809	.6210	.6067
	Grammar: Independent	.0781	–	–
	Usage: Independent	.0573	–	–
	Mechanics: Independent	.0542	–	–
	Vocabulary sophistication: Independent	.0476	–	–
	Word complexity: Integrated	.0223	–	–
	Syntactic variety: Independent	.0291	–	–
	Development: Independent	.1450	–	–
	Organization: Independent	.1635	–	–
	Collocation-preposition: Independent	.0193	–	–
	Grammar: Integrated	.0414	–	–
	Usage: Integrated	.0176	–	–
	Mechanics: Integrated	.0201	–	–
	Vocabulary sophistication: Integrated	.0341	–	–
	Word complexity: Integrated	−.0086	–	–
	Syntactic variety: Integrated	.0350	–	–
	Development: Integrated	.0881	–	–
	Organization: Integrated	.0729	–	–
	Collocation-preposition: Integrated	.0093	–	–

Note. With $\mathbf{M}_1$, the predictors are the first human score and nine e-rater feature scores for each prompt. With $\mathbf{M}_2$, the predictors only include the first human score for each prompt. With $\mathbf{M}_3$, the predictor is the average of two human scores for each prompt.

Table 3 shows that Praxis Writing exhibits very different behavior than that seen in the other two tests. The estimated standardized regression coefficient for human rating with $\mathbf{M}_1$ (.2069) is substantially smaller than the corresponding values with $\mathbf{M}_2$ (.8449) and $\mathbf{M}_3$ (.9217). Again, organization and development features appear to have the greatest importance with $\mathbf{M}_1$, in which model the importance indeed exceeds that of human holistic scores. The reasons for this difference in behavior for Praxis Writing merit comparison of the human-scoring processes of the three tests; however, this comparison is beyond the scope of this report.

Results for Assessment Accuracy (Research Question 1)

Assessment accuracy was evaluated by use of repeater samples. MDIA was used to compensate for the unrepresentative sampling of repeaters. In this case, the estimates of PRMSE statistics, for the two weighting choices for TOEFL Writing and GRE Writing and one weighting case for Praxis Writing, imply how reliably the tests measure the construct of interest.

Table 2 Estimated Standardized Regression Coefficients for Scoring Accuracy: GRE Writing

Weight	Variable	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_3$
Equal	Human rating: Issue	.2450	.5008	.5057
	Human rating: Argument	.3211	.5221	.5301
	Grammar: Issue	.0394	–	–
	Usage: Issue	.0637	–	–
	Mechanics: Issue	.0528	–	–
	Vocabulary sophistication: Issue	.0333	–	–
	Word complexity: Issue	.0386	–	–
	Syntactic variety: Issue	.0359	–	–
	Development: Issue	.1671	–	–
	Organization: Issue	.1795	–	–
	Collocation-preposition: Issue	.0302	–	–
	Grammar: Argument	.0344	–	–
	Usage: Argument	.0528	–	–
	Mechanics: Argument	.0230	–	–
	Vocabulary sophistication: Argument	.0213	–	–
	Word complexity: Argument	.0117	–	–
	Syntactic variety: Argument	.0463	–	–
	Development: Argument	.1405	–	–
	Organization: Argument	.1650	–	–
	Collocation-preposition: Argument	.0239	–	–
Unequal	Human rating: Issue	.2479	.5072	.5146
	Human rating: Argument	.3162	.5157	.5213
	Grammar: Issue	.0400	–	–
	Usage: Issue	.0645	–	–
	Mechanics: Issue	.0537	–	–
	Vocabulary sophistication: Issue	.0337	–	–
	Word complexity: Issue	.0398	–	–
	Syntactic variety: Issue	.0360	–	–
	Development: Issue	.1754	–	–
	Organization: Issue	.1887	–	–
	Collocation-preposition: Issue	.0302	–	–
	Grammar: Argument	.0344	–	–
	Usage: Argument	.0530	–	–
	Mechanics: Argument	.0227	–	–
	Vocabulary sophistication: Argument	.0209	–	–
	Word complexity: Argument	.0110	–	–
	Syntactic variety: Argument	.0458	–	–
	Development: Argument	.1339	–	–
	Organization: Argument	.1569	–	–
	Collocation-preposition: Argument	.0238	–	–

Note. With $\mathbf{M}_1$, the predictors are the first human score and nine e-rater feature scores for each prompt. With $\mathbf{M}_2$, the predictors only include the first human score for each prompt. With $\mathbf{M}_3$, the predictor is the average of two human scores for each prompt.

Generally, the estimates of PRMSE in case of assessment accuracy are consistently lower than the estimates in case of scoring accuracy. This phenomenon has been observed in earlier studies (Haberman et al., 2015; Haberman & Yao, 2015). The underlying reasons are that scoring accuracy only involves the examination of the scoring qualities of human judgment on the present test; however, assessment accuracy reflects not only variations in examinee responses to different prompts but also variations in examinee learning between assessments taken at different times. As a consequence, it is common to have lower PRMSE values in case of assessment accuracy when compared to PRMSE measures in case of scoring accuracy.

Despite the lower estimates, the overall patterns of the estimated PRMSE statistics in the case of assessment accuracy are similar to those in the case of scoring accuracy. With $\mathbf{M}_2$, the estimated PRMSE measures for all three test are the lowest (i.e., .7077 with equal weighting and .7064 with unequal weighting for TOEFL, .7632 with equal weighting and .7632 with unequal weighting for GRE, and .4410 for Praxis). The employment of nine e-rater feature scores ($\mathbf{M}_3$), compared to $\mathbf{M}_2$, leads to a substantial increase in the estimated PRMSE measures for all three tests, that is, .8065 with equal weighting and .8135 with unequal weighting for TOEFL, .8564 with equal weighting and .8567 with unequal

Table 3 Estimated Standardized Regression Coefficients for Scoring Accuracy: Praxis Writing

Variable	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_3$
Human rating	.2069	.8449	.9217
Grammar	.0551	–	–
Usage	.0397	–	–
Mechanics	.1097	–	–
Vocabulary sophistication	.0404	–	–
Word complexity	.1578	–	–
Syntactic variety	.0401	–	–
Development	.6268	–	–
Organization	.8363	–	–
Collocation-preposition	.0160	–	–

Note. With $\mathbf{M}_1$, the predictors are the first human score and nine e-rater feature scores for each prompt. With $\mathbf{M}_2$, the predictors only include the first human score for each prompt. With $\mathbf{M}_3$, the predictor is the average of two human scores for each prompt.

weighting for GRE, and .6002 for Praxis. In addition, two human raters per prompt ($\mathbf{M}_3$) yields notably better results than those for one human rater per prompt ($\mathbf{M}_2$). The estimates of PRMSEs with $\mathbf{M}_3$ are .8289 with equal weighting and .8275 with unequal weighting for TOEFL, .8657 with equal weighting and .8657 with unequal weighting for GRE, and .6121 for Praxis. Using e-rater features and one human rating per prompt ($\mathbf{M}_1$) was a bit less effective than double human scoring ($\mathbf{M}_3$) in all cases, although much more effective than just single human scoring. The Praxis situation was strikingly different for assessment accuracy than for scoring accuracy, for double human scoring ($\mathbf{M}_3$) becomes more effective than e-rater features and single human scoring ($\mathbf{M}_1$). The relatively low PRMSEs for Praxis reflect use of only one prompt.

Lastly, the results of the estimated PRMSE values suggest that, for both tests, the choice of weighting procedure has little to no impact on the estimates of PRMSE. Similar to the scoring accuracy results, the weights c_k are rather even for GRE Writing: .5103 and .4897 for issue and argument prompts, respectively. Comparable values for TOEFL Writing are .5998 (independent) and .4002 (integrated).

For detailed information, Table 4 gives the estimated standardized regression coefficients for TOEFL Writing with three matrices $\mathbf{M}$ for both equal and unequal weighting of prompts.

Comparable results for GRE Writing are presented in Table 5. The general pattern of the relative contributions of each predictor for three matrices $\mathbf{M}$, regardless of whether equal or unequal weights are employed, is highly similar to the results for scoring accuracy. That is, for both tests, small differences are found between $\mathbf{M}_2$ and $\mathbf{M}_3$ in terms of the relative weights for the human ratings for the two prompts, adding e-rater features in $\mathbf{M}_1$ substantially reduces the weights for human ratings, and organization and development receive the highest weights among all e-rater features. As in the case of scoring accuracy, the human ratings on the less complex writing prompt (i.e., independent for TOEFL Writing and issue for GRE Writing) exhibit a larger decrease in weight for $\mathbf{M}_1$ than do the human ratings on the other prompt.

Comparison of standardized regression coefficients for Praxis Writing in Tables 3 – 6 indicates roughly comparable relative contributions of the single human score and the essay features despite the somewhat different results for comparison of $\mathbf{M}_1$ and $\mathbf{M}_2$ for scoring accuracy and for assessment accuracy.

Subgroup Analyses Based on Penalized Best Linear Predictor Method (Research Question 2)

To fully answer the second research question, this section provides the results based on the PBLP method in case of assessment accuracy and $\mathbf{M} = \mathbf{M}_1$, to complement the information in Yao et al. (2019) that presented the results of scoring accuracy only for illustration. As in Haberman et al. (2015), both definitions of scoring accuracy and assessment accuracy are meaningful, hence this section mainly presents the results of assessment accuracy to supplement the subgroup analyses described in Yao et al. (2019). Similar to Yao et al., we still evaluate the effectiveness of the penalty function by two measures: the comparison of the estimate of the variance $\sigma^2\left(E\left(r_{\mathbf{M}_1 d}|G\right)\right)$ for $d > 1$ and $d = 1$ (no penalty) and the comparison of the estimated differences between $E\left(\widehat{\xi}_{\mathbf{M}_1 d}|G = h\right)$ and $E(O|G = h)$ for $1 \le h \le H$ and for both $d > 1$ and $d = 1$. In Appendix C, Tables C1 – C8 summarize the results. In addition, Figures 1 and 2 graphically illustrate the results shown in Tables C4 and C5.

Table 4 Estimated Standardized Regression Coefficients for Assessment Accuracy: TOEFL Writing

Weight	Variable	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_3$
Equal	Human rating: Independent	.1900	.4296	.4496
	Human rating: Integrated	.3717	.5367	.5545
	Grammar: Independent	.0912	–	–
	Usage: Independent	.0701	–	–
	Mechanics: Independent	.0559	–	–
	Vocabulary level: Independent	.0563	–	–
	Word complexity: Independent	.0083	–	–
	Syntactic variety: Independent	.0603	–	–
	Development: Independent	.1347	–	–
	Organization: Independent	.1445	–	–
	Collocation-preposition: Independent	.0253	–	–
	Grammar: Integrated	.0814	–	–
	Usage: Integrated	.0443	–	–
	Mechanics: Integrated	.0317	–	–
	Vocabulary level: Integrated	.0580	–	–
	Word complexity: Independent	.0121	–	–
	Syntactic variety: Integrated	.0264	–	–
	Development: Integrated	.1468	–	–
	Organization: Integrated	.1461	–	–
	Collocation-preposition: Integrated	.0006	–	–
Unequal	Human rating: Independent	.2009	.6494	.6796
	Human rating: Integrated	.3459	.8113	.8382
	Grammar: Independent	.0930	–	–
	Usage: Independent	.0764	–	–
	Mechanics: Independent	.0613	–	–
	Vocabulary level: Independent	.0573	–	–
	Word complexity: Integrated	.0094	–	–
	Syntactic variety: Independent	.0620	–	–
	Development: Independent	.1422	–	–
	Organization: Independent	.1563	–	–
	Collocation-preposition: Independent	.0261	–	–
	Grammar: Integrated	.0855	–	–
	Usage: Integrated	.0478	–	–
	Mechanics: Integrated	.0350	–	–
	Vocabulary level: Integrated	.0588	–	–
	Word complexity: Integrated	.0103	–	–
	Syntactic variety: Integrated	.0311	–	–
	Development: Integrated	.1398	–	–
	Organization: Integrated	.1341	–	–
	Collocation-preposition: Integrated	.0019	–	–

Note. With $\mathbf{M}_1$, the predictors are the first human score and nine e-rater feature scores for each prompt. With $\mathbf{M}_2$, the predictor is only the first human score for each prompt. With $\mathbf{M}_3$, the average of two human scores for each prompt is the predictor.

Tables C1 – C3 record several measures to evaluate the effectiveness of the PBLP method, in terms of the estimated variance of the observed composite score $\sigma^2(O)$, the estimated MSE for $\mathbf{M}_1$ as $\mathrm{MSE}_{\mathbf{M}_1 d}$ based on penalty multiplier $d-1$, the estimated variance of the conditional residual expectation $\sigma^2\left(E\left(r_{\mathbf{M}_1 d}|G\right)\right)$, and the estimated PRMSE for $\mathbf{M}_1$, $\rho^2_{\mathbf{M}_1 d}$. Note that when $d=1$, there is no penalty, in which case, PBLP reduces to the regular BLP. Therefore the $\rho^2_{\mathbf{M}_1 d}$ measures when $d=1$, for each assessment studied, in Tables C1 – C3 are the same as the estimated PRMSE measures when $\mathbf{M}_1$ is assessed.

From Tables C1 – C3, we observe that as the penalty strength d increases, the estimated values $\sigma^2\left(E\left(r_{\mathbf{M}_1 d}|G\right)\right)$ go down substantially for all three assessments. Meanwhile, the estimated values of the PRMSE $\rho^2_{\mathbf{M}_1 d}$ for an assessment dropped. However, the PRMSE $\rho^2_{\mathbf{M}_1 d}$ remains above .8000 when the penalty multiplier d goes up to 4 or 5 for TOEFL Writing, and remains beyond .7500 even when d reaches 100. For GRE Writing, the PRMSE measure only drops by about 0.03 points when the penalty multiplier d increases from 1 to 100, whereas the estimated values $\sigma^2\left(E\left(r_{\mathbf{M}_1 d}|G\right)\right)$ go down considerably. A similar compensatory pattern is observed for Praxis Writing. It is also evident that the two sets of weighting, either

Table 5 Estimated Standardized Regression Coefficients for Assessment Accuracy: GRE Writing

Weight	Variable	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_3$
Equal	Human rating: Issue	.2283	.4993	.5116
	Human rating: Argument	.2335	.4683	.4814
	Grammar: Issue	.0294	–	–
	Usage: Issue	.0936	–	–
	Mechanics: Issue	.0715	–	–
	Vocabulary level: Issue	.0623	–	–
	Word complexity: Issue	.0139	–	–
	Syntactic variety: Issue	.0551	–	–
	Development: Issue	.1006	–	–
	Organization: Issue	.1054	–	–
	Collocation-preposition: Issue	.0614	–	–
	Grammar: Argument	.0680	–	–
	Usage: Argument	.0729	–	–
	Mechanics: Argument	.0311	–	–
	Vocabulary level: Argument	.0271	–	–
	Word complexity: Argument	.0247	–	–
	Syntactic variety: Argument	.0639	–	–
	Development: Argument	.1554	–	–
	Organization: Argument	.1592	–	–
	Collocation-preposition: Argument	.0175	–	–
Unequal	Human rating: Issue	.2282	.5004	.5134
	Human rating: Argument	.2325	.4672	.4796
	Grammar: Issue	.0294	–	–
	Usage: Issue	.0942	–	–
	Mechanics: Issue	.0713	–	–
	Vocabulary level: Issue	.0620	–	–
	Word complexity: Issue	.0144	–	–
	Syntactic variety: Issue	.0549	–	–
	Development: Issue	.1028	–	–
	Organization: Issue	.1077	–	–
	Collocation-preposition: Issue	.0614	–	–
	Grammar: Argument	.0681	–	–
	Usage: Argument	.0732	–	–
	Mechanics: Argument	.0316	–	–
	Vocabulary level: Argument	.0272	–	–
	Word complexity: Argument	.0243	–	–
	Syntactic variety: Argument	.0639	–	–
	Development: Argument	.1547	–	–
	Organization: Argument	.1578	–	–
	Collocation-preposition: Argument	.0171	–	–

Note. With $\mathbf{M}_1$, the predictors are the first human score and nine e-rater feature scores for each prompt. With $\mathbf{M}_2$, the predictor is only the first human score for each prompt. With $\mathbf{M}_3$, the average of two human scores for each prompt is the predictor.

equal weighting or unequal weighting, result in similar trends for TOEFL and GRE tests. More importantly, compared to the results of scoring accuracy reported in Yao et al. (2019), the overall patterns in case of assessment accuracy are quite consistent, except for the lower value of $\rho^2_{\mathbf{M}_1 d}$ and more drop as d increases. On the whole, similar to Yao et al., this supplementary information based on assessment accuracy well demonstrates the trade-off of the PBLP approach between the accuracy and the fairness as d increases, meanwhile keeping the overall accuracy of BLP. It is also noticeable that when $d = 4$ or $d = 5$, the results start becoming stable, whereas when d turns to two digits, the changes are quite small within most of subgroups.

Tables C4 – C8 present the results of assessment accuracy for TOEFL Writing, GRE Writing, and Praxis Writing, respectively, for comparison of the estimated values of $E\left(\widehat{\zeta}_{\mathbf{M}_1 d}|G = h\right)$ and the estimated values of $E(O|G = h)$ to assess the impact on subgroup differences when adding computer-generated essay features in the prediction model. The subgroups are not explicitly identified in accordance with policy of the data sources, so we name the subgroups by numerical numbers as in Yao et al. (2019), although the sample sizes of each subgroup are recorded in detail. Besides, we provide Figures 1

Table 6 Estimated Standardized Regression Coefficients for Assessment Accuracy: Praxis Writing

Variable	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_3$
Human rating	.1461	.6641	.7842
Grammar	.0748	–	–
Usage	.0551	–	–
Mechanics	.1657	–	–
Vocabulary level	.0674	–	–
Word complexity	.1357	–	–
Syntactic variety	.0521	–	–
Development	.5273	–	–
Organization	.6166	–	–
Collocation-preposition	−.0160	–	–

Note. With $\mathbf{M}_1$, the predictors are the first human score and nine e-rater feature scores for each prompt. With $\mathbf{M}_2$, the predictor is only the first human score for each prompt. With $\mathbf{M}_3$, the average of two human scores for each prompt is the predictor.

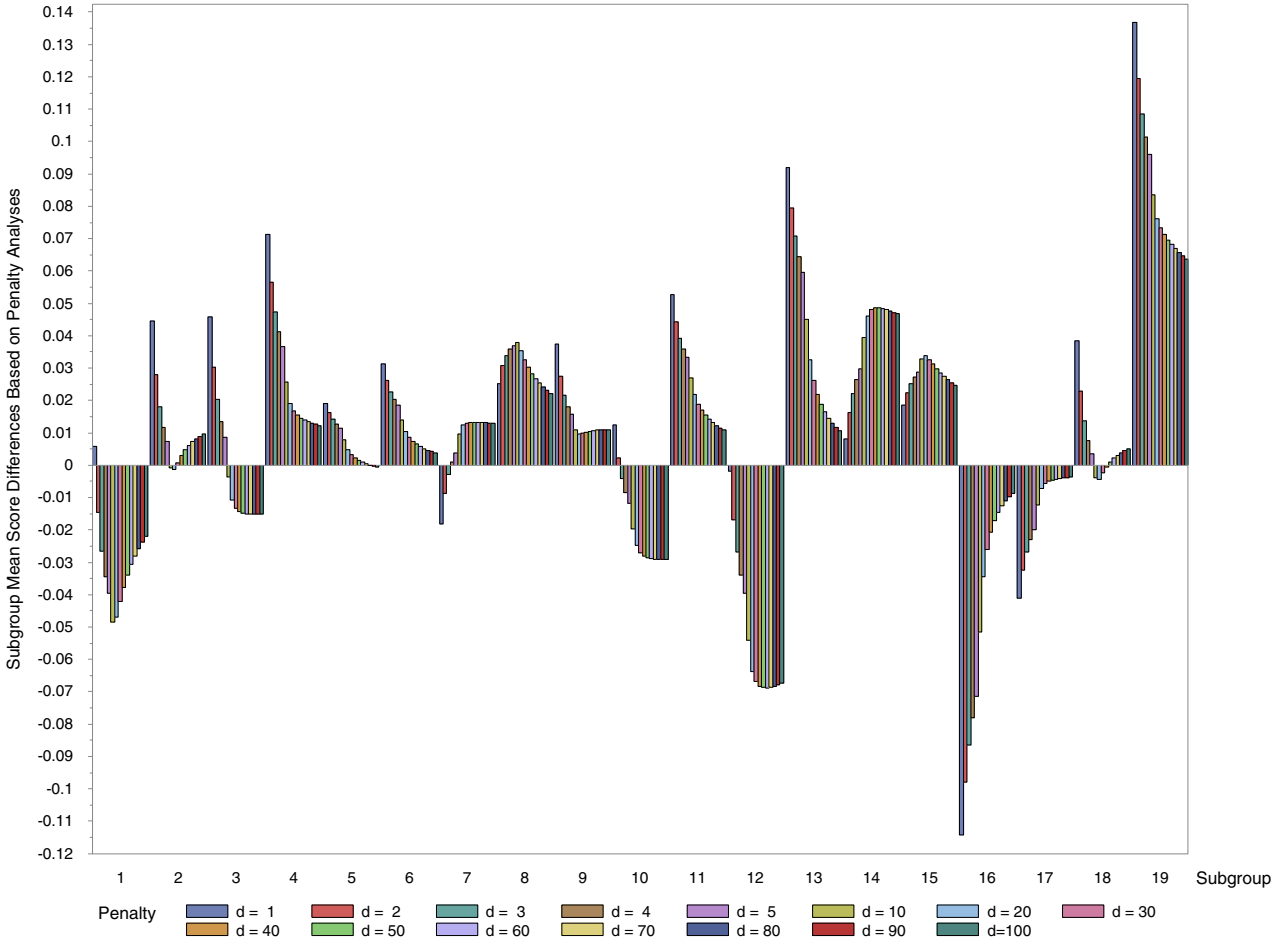

Figure 1 Estimated mean differences for assessment accuracy: TOEFL Writing, equal weights.

and 2 to show the results in Tables C4 and C5 by graphic demonstration. Similar to Yao et al. (2019), the results in Tables C4 and C5 show that the magnitude of the absolute mean differences between $E\left(\hat{\zeta}_{\mathbf{M}_1 d}|G=h\right)$ and $E(O|G=h)$ for the majority of the subgroups tends to reduce as the penalty parameter d goes up. The effectiveness of the penalty function has similar impact for both GRE Writing and Praxis Writing, as seen in Tables C6 – C8. Consistent with previous results, the two ways of weighting strategies yield similar results in TOEFL Writing and GRE Writing. It is also verified that this part of the results becomes stable when $d=4$ or $d=5$; especially when d turns to two digits, the changes in the magnitude

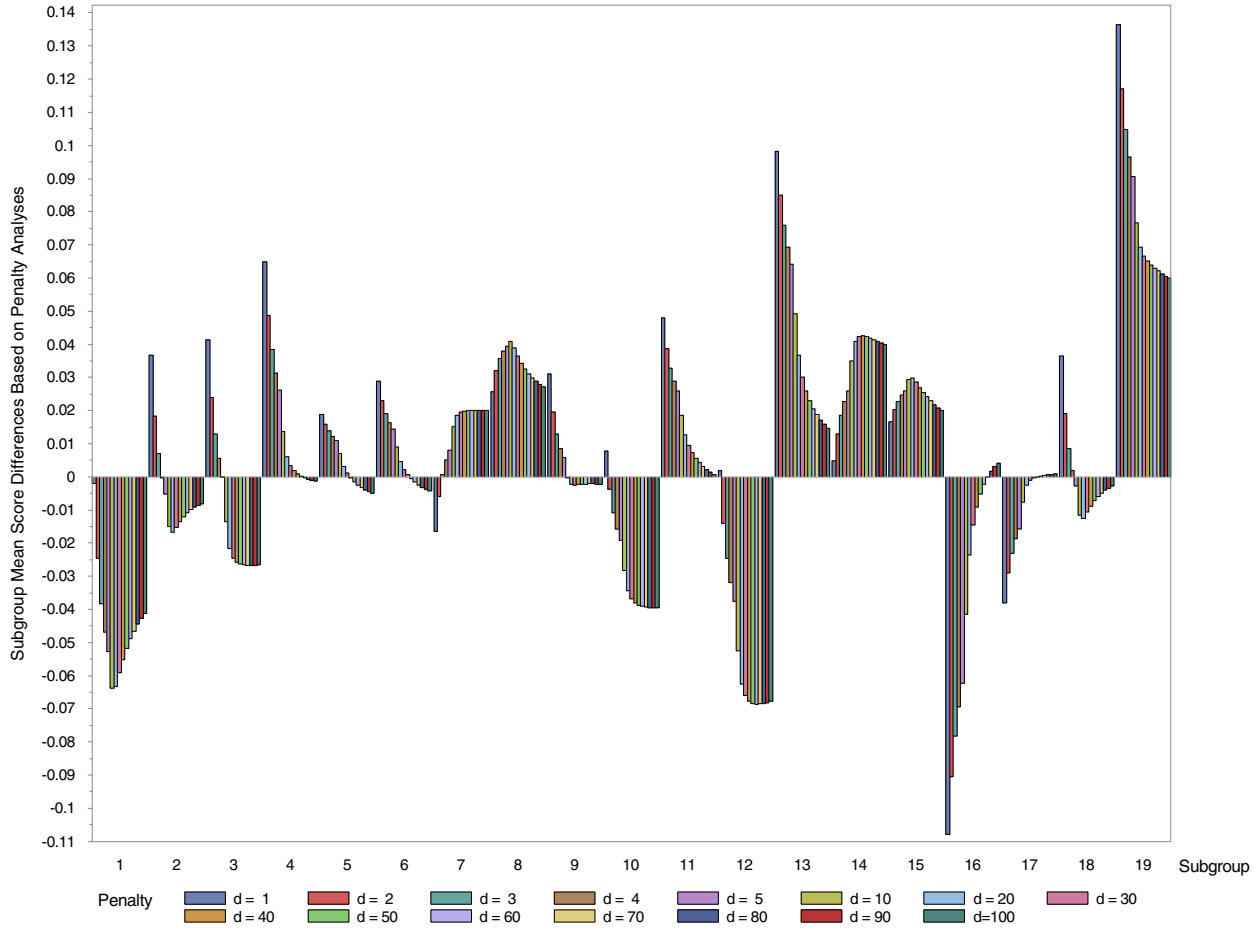

Figure 2 Estimated mean differences for assessment accuracy: TOEFL Writing, unequal weights.

are quite small within most subgroups. In a more direct way, the figures illustrate that the initial large mean score difference (when $d = 1$ with no penalty at the beginning) among most subgroups tends to diminish substantially as d increases. Moreover, this effect is most clear in subgroups with relative large size that have relatively large initial mean score differences. However, compared to the results of scoring accuracy in Yao et al., the magnitude of mean score differences based on assessment accuracy are comparatively larger, as seen in Tables C4 – C8.

Based on the results of PBLP in the case of scoring accuracy in Yao et al. (2019) and the results of PBLP in the case of assessment accuracy reported in this section, the proposed PBLP method is shown to effectively treat the subgroup biases, for all three testing programs studied. Lastly, it is worthwhile concluding that there is an obvious trade-off between the accuracy and the fairness of the PBLP method due to increasing d, for which we do not have an universal rule to determine the optimal d and the selection fully depends on balance of accuracy and fairness.

Discussion

This report, in great detail, describes the methodology of the BLP approach developed in Haberman et al. (2015) and a modified version of the PBLP approach proposed in Yao et al. (2019) to treat subgroup biases. This report gives full accounts of the results of the applications of these methodologies to three ETS operational testing programs that use automated scoring more extensively than what was available in Haberman et al. (2015). This report can further serve as supplementary material to Yao et al. (2019). In the case of BLP, compared to the study of Haberman et al. (2015), the BLP approach is flexible in terms of the number of writing prompts in an assessment. For the data studied, two prompts were used in TOEFL Writing and GRE Writing, while one essay prompt was used in Praxis Writing. In the case of PBLP, in addition to the results of applications described in Haberman et al. (2015), we added another set of results in the

case of assessment accuracy for the subgroup analyses for demonstrating the full picture of the effectiveness of PBLP in improving the population invariance. To be more specific, this report mainly addressed two research questions. One question related to the performance of the BLP model using human and machine features compared to all-human scoring in terms of scoring and assessment accuracy, and the second question concerned the reduction of subgroup effects with PBLP.

We compared three alternative predictor sets. For the TOEFL Writing and GRE Writing, analysis also varied in two ways according to the weighting strategy for the two essay prompts. The models were evaluated and compared based on scoring and assessment accuracy (Research Question 1) and deviation from population invariance (Research Question 2). For scoring accuracy, agreement samples were used for the estimation of the variance of the measurement errors. Assessment accuracy was studied by the use of repeater data weighted by MDIA, as in Haberman (1984).

The results, consistently for all three testing programs, revealed that using two human scores per prompt ($\mathbf{M}_3$) produced much greater scoring and assessment accuracy than using only one human score per prompt ($\mathbf{M}_2$). In a similar manner, using automated essay features and one human rating for each essay prompt ($\mathbf{M}_1$) substantially outperformed using only one human score ($\mathbf{M}_2$), in terms of both scoring accuracy and assessment accuracy, for all three tests. The picture for comparison of computer-generated essay features and one human score ($\mathbf{M}_1$) and double human scoring ($\mathbf{M}_3$) was somewhat more complex. For TOEFL Writing, double human scoring dominated for both scoring and assessment accuracy. The dominance for scoring accuracy increased as steps were taken to reduce subgroup discrepancies. For GRE Writing, results for $\mathbf{M}_1$ and $\mathbf{M}_3$ were rather close, although double human scoring dominated for both scoring accuracy and assessment accuracy. Domination for scoring accuracy was increased as more effort was made to treat subgroup effects. For Praxis Writing, double human scoring was dominated strongly by the combination with $\mathbf{M}_1$ of computer-generated features and single human scoring; however, double human scoring had a small advantage for assessment accuracy.

This report additionally supplements Yao et al. (2019) by providing the results of subgroup analyses in the case of assessment accuracy. Similar to the results of scoring accuracy, the between-group mean variance reduces when the penalty parameter increases, which suggests better population invariance. On the other hand, the mean differences between scaled predicted writing true scores on the subgroups resulting from PBLP models and the observed composite score based on the base model (only human holistic scores without any penalty) tended to diminish as the penalty parameter increased. The phenomenon is most evident for the subgroups that had initial large mean differences prior to the application of the penalty function. Thus it is well verified that the proposed PBLP method effectively treated subgroup biases related to the interaction of the machine feature scores and the demographic characteristics of the examinees. Although the overall patterns of the results based on assessment accuracy are similar to those results based on scoring accuracy shown in Yao et al., some difference still exists, such as lower value of ρ^2 and more drop of ρ^2 as d rises. The added information in this report discloses the full picture of how the PBLP method treats the subgroup biases without losing overall accuracy, either in the case of scoring accuracy or in the case of assessment accuracy.

Regression coefficients for scoring accuracy are not the same as those in assessment accuracy. Space limitations prevent a thorough examination of this issue here, but it should be noted that results from use of scoring accuracy as a criterion were evaluated in terms of the criterion of assessment accuracy and results from use of assessment accuracy as a criterion were evaluated in terms of the criterion of scoring accuracy. For the data under study, results were quite robust, except that rescaling is appropriate due to the difference in PRMSE for scoring accuracy and for assessment accuracy. No guarantee exists that all data will behave in this fashion.

However, some caveats should be noted. In traditional assessments, Cronbach's alpha is commonly used to evaluate assessment reliability without use of repeater data. This approach requires multiple prompts, so it certainly does not apply to Praxis Writing. Even for TOEFL Writing and GRE Writing, two prompts is not entirely satisfactory. Nonetheless, this traditional approach is worth consideration when the number J of items is larger. Use of Cronbach's alpha is even more challenging for TOEFL Writing, for there are significant issues due to quite different constructs measured and due to unequal variances of responses for the two prompts. There is a further issue not discussed in the report: Rater reliability for human scorers is much better for the integrated prompt when compared to the independent prompt. As in all approaches with repeater data, caution must always be exercised in interpreting results. Even with MDIA, concern must exist that sampling bias has not been entirely corrected by weighting. This situation is most serious when sampling bias is largest.

On the whole, the results in this study show the added value of using machine features to predict composite writing true scores and effective use of a penalty function to achieve greater population invariance. Finally, it should be noted that the suggested scoring methods are not limited to writing assessments. They have general practical implications for testing programs that intend to use automated scoring capabilities or score augmentation for score reporting, especially whenever subgroup biases need to be treated.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology*, *Learning, and Assessment*, *4*(3), 1 – 29. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Attali, Y., Burstein, J., & Andreyev, S. (2003). *E-rater version 2.0: Combining writing analysis feedback with automated essay scoring*. Unpublished manuscript.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, *25* (3), 27 – 36. https://doi.org/10.1609/aimag.v25i3.1774

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37* , 281 – 306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12* , 971 – 988. https://doi.org/10.1214/aos/1176346715

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33* , 204 – 229. https://doi.org/10.3102/1076998607302636

Haberman, S. J., & Qian, J. (2007). Linear prediction of a true score from a direct estimate and several derived estimates. *Journal of Educational and Behavioral Statistics*, *32* , 6 – 23. https://doi.org/10.3102/1076998606298036

Haberman, S. J., & Sinharay, S. (2010). The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics*, *35* , 586 – 602. https://doi.org/10.3102/1076998610375839

Haberman, S. J., & Sinharay, S. (2013). Does subgroup membership information lead to better estimation of true subscores? *British Journal of Mathematical and Statistical Psychology*, *66* , 452 – 469. https://doi.org/10.1111/j.2044-8317.2012.02061.x

Haberman, S. J., & Yao, L. (2015). Repeater analysis for combining information from different assessments. *Journal of Educational Measurement*, *52* , 223 – 251. https://doi.org/10.1111/jedm.12075

Haberman, S. J., Yao, L., & Sinharay, S. (2015). Prediction of true test scores from observed item scores and ancillary data. *British Journal of Mathematical and Statistical Psychology*, *68*, 363 – 385. https://doi.org/10.1111/bmsp.12052

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley. https://doi.org/10.1002/9780470316436

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, *37* , 113 – 140. https://doi.org/10.1111/j.1745-3984.2000.tb01079.x

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Swygert, K. A., & Thissen, D. (2001). Augmented scores: "Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343 – 387). Mahwah, NJ: Erlbaum.

Yao, L., Haberman, S., & Zhang, M. (2019). Penalized best linear prediction of test true scores. *Psychometrika*, *84* , 186 – 211. https://doi.org/10.1007/s11336-018-9636-7

Appendix A: Modification of Formulas for Estimation of Best Linear Predictor and Penalized Best Linear Predictor

We fail to observe $X_{i(J+j)}$, $1 \le j \le J$, for most observations i, $1 \le i \le n$. Instead, for prompt j, $1 \le j \le J$, a subsample U_j with $u_j > 0$ members is obtained from the observations $1 \le i \le n$. In typical cases, u_j is much smaller than n. Only for observations i in the subsample U_j is the second human rating $X_{i(J+j)}$ observed. Because $\sigma^2(X_{J+j}) = \mathrm{Cov}(X_{J+j}, X_{J+j}) = \mathrm{Cov}(X_j, X_j) = \sigma^2(X_j)$, the variance $\sigma^2(X_j - X_{J+j}) = 2[\mathrm{Cov}(X_j, X_j) - \mathrm{Cov}(X_{J+j}, X_j)]$ of the difference $X_j - X_{J+j}$ has the unbiased estimate

$$\overline{\sigma}^2\left(X_j - X_{J+j}\right) = \left(u_j\right)^{-1} \sum_{i \in U_j} \left(X_{ij} - X_{i(J+j)}\right)^2. \tag{A1}$$

Because $E(X_j) = E(X_{J+j})$ for $1 \le j \le J$, and $\mathrm{Cov}(X_j, X_k) = \mathrm{Cov}(X_{J+j}, X_k)$ for $1 \le j \le J$ and $1 \le k \le K$ such that k is neither j nor $J+j$, the following practices are adopted. Let $\widetilde{X}_i$, $1 \le i \le n$, be the vector with K elements $\widetilde{X}_{ik}$, $1 \le k \le K$, such that $\widetilde{X}_{ik} = X_{ik}$ for $1 \le k \le J$ and $2J < k \le K$ and $\widetilde{X}_{ik} = X_{i(k-J)}$ for $J < k \le 2J$. Then $\overline{E}(\mathbf{X})$ is replaced by $\overline{E}\left(\widetilde{\mathbf{X}}\right)$, so that $\overline{E}\left(\widetilde{\mathbf{X}}\right)$ has expectation $E(\mathbf{X})$. Let $\overline{\mathbf{D}}(\mathbf{X})$ be the K by K matrix with elements $\overline{D}_{kk'}(\mathbf{X})$ for $1 \le k \le K$ and column k', $1 \le k' \le K$, such that $\overline{D}_{kk'}(\mathbf{X}) = 0$ if k or k' is greater than $2J$ or $|k-k'|$ is not J and $\overline{D}_{j(J+j)}(\mathbf{X}) = \overline{D}_{(J+j)j}(\mathbf{X}) = \frac{1}{2}\overline{\sigma}^2\left(X_j - X_{J+j}\right)$ for $1 \le j \le J$. Then $\overline{\mathrm{Cov}}(\mathbf{X})$ is replaced by $\overline{\mathrm{Cov}}_m(\mathbf{X}) = \overline{\mathrm{Cov}}\left(\widetilde{\mathbf{X}}\right) - \overline{\mathbf{D}}(\mathbf{X})$. This substitution is made because $\overline{\mathrm{Cov}}_m(\mathbf{X})$ has expectation $\mathrm{Cov}(\mathbf{X})$.

Appendix B: Estimation of Minimum Discriminant Information Adjustment Weights

To describe the procedure, for $1 \le i \le n$, let S_i be 1 if $R(t(i)) > 1$ and $i = i_*(1, t(i))$. If $S_i = 1$, let $i_2(i)$ be $i_*(2, t(i))$, the observation for the second time test taker $t(i)$ takes the assessment. Let U be the set of examinees i with $S_i = 1$, and let U have n_U elements. Assume that the expectation of n_U/n approaches a positive constant as the sample size n becomes large. The sample weights w_i, i in U, are selected so that the weighted sample of $\left(\mathbf{X}_i, \mathbf{X}_{i_2(i)}\right)$, i in U, has a distribution that shares many features that should be associated with the distribution of $(\mathbf{X}, \mathbf{X}_*)$, even though the repeater observations are not representative. For this purpose, Q-dimensional random variables $\mathbf{Y}_i$, i in U, and $\mathbf{Z}_i$, $1 \le i \le n$, are defined such that $\mathbf{Y}_i$ is a function of $\left(\mathbf{X}_i, \mathbf{X}_{i_2(i)}\right)$ and $\mathbf{Z}_i$ is a function of $\mathbf{X}_i$. The weights w_i, i in U, are selected to minimize the sample discriminant information $-n_U^{-1}\sum_{i\in U} w_i \log\left(w_i\right)$ subject to the constraints

$$n_U^{-1}\sum_{i\in U} w_i = 1 \tag{B1}$$

and

$$n_U^{-1}\sum_{i\in U} w_i \mathbf{Y}_i = \overline{E}(\mathbf{Z}) = n^{-1}\sum_{i=1}^{n} \mathbf{Z}_i. \tag{B2}$$

In a variation on Haberman et al. (2015), let the vector consisting of the initial K elements of $\mathbf{Y}_i$ be $\mathbf{X}_i$, and let the same condition apply to $\mathbf{Z}_i$. Let the next $K(K+1)/2$ elements of $\mathbf{Y}_i$ and the next $K(K+1)/2$ elements of $\mathbf{Z}_i$ be $X_{ik}X_{ik'}$, $1 \le k \le k' \le K$. Then let the next $K(K+1)/2$ elements of $\mathbf{Y}_i$ be $X_{i_2(i)k}X_{i_2(i)k'*}$ for $1 \le k \le k' \le K$, and let the next $K(K+1)/2$ elements of $\mathbf{Z}_i$ be $X_{ik}X_{ik'}$ for $1 \le k \le k' \le K$. Next, let the next $K(K-1)/2$ elements of $\mathbf{Y}_i$ be $X_{ik}X_{i_2(i)k'} - X_{ik'}X_{i_2(i)k}$ for $1 \le k < k' \le K$, and let the corresponding elements of $\mathbf{Z}_i$ be 0. Let the final $H-1$ elements of $\mathbf{Y}_i$ and the final $H-1$ elements of $\mathbf{Z}_i$ be $\delta_h(G_i)$, $1 \le h \le H-1$, where $\delta_h(h')$ is 1 for real $h = h'$, and 0 otherwise. The resulting weights w_i lead to the estimate

$$\overline{\mathrm{Cov}}(\boldsymbol{\tau}) = \overline{\mathrm{Cov}}\left(\mathbf{X}, \mathbf{X}_*\right) = n_U^{-1}\sum_{i\in U} w_i\left[\mathbf{X}_i - \overline{E}(\mathbf{X})\right]\left[\mathbf{X}_{i_2(i)} - \overline{E}(\mathbf{X})\right]' \tag{B3}$$

of $\mathrm{Cov}(\boldsymbol{\tau})$. The constraints of MDIA imply that $\overline{\mathrm{Cov}}(\boldsymbol{\tau})$ is symmetric:

$$n_U^{-1}\sum_{i\in U} w_i\mathbf{X}_i = n_U^{-1}\sum_{i\in U} w_i\mathbf{X}_{i_2(i)} = \overline{E}(\mathbf{X}), \tag{B4}$$

$$\begin{aligned} n_U^{-1}\sum_{i\in U} w_i\left[\mathbf{X}_i - \overline{E}(\mathbf{X})\right]\left[\overline{X}_i - \overline{E}(\mathbf{X})\right]' &= n_U^{-1}\sum_{i\in U} w_i\left[\mathbf{X}_{i_2(i)} - \overline{E}(\mathbf{X})\right]\left[\overline{X}_{i_2(i)} - \overline{E}(\mathbf{X})\right]' \\ &= \overline{\mathrm{Cov}}(\mathbf{X}), \end{aligned} \tag{B5}$$

and

$$n_U^{-1}\sum_{i\in U} w_i\delta_h\left(G_i\right) = \overline{p}_G(h), 1 \le h \le H. \tag{B6}$$

When agreement samples are involved, MDIA uses $\widetilde{\mathbf{X}}_i$, $1 \le i \le n$, instead of $\mathbf{X}_i$, $1 \le i \le n$, and the definition of $\overline{\mathrm{Cov}}(\boldsymbol{\tau})$ uses $\widetilde{\mathbf{X}}_i$ instead of $\mathbf{X}_i$

Appendix C: Additional Tables

Table C1 Estimated Parameters for Penalized Best Linear Predictor for TOEFL Writing: Assessment Accuracy

Weight	d	$\sigma^2(O)$	$MSE_{M_1 d}$	$\rho^2_{M_1 d}$	$\sigma^2\left(E\left(r_{M_{1d}}\|G\right)\right)$
Equal	1	0.4761	0.0921	0.8065	0.0038
	2	0.4761	0.0926	0.8055	0.0027
	3	0.4761	0.0935	0.8036	0.0021
	4	0.4761	0.0945	0.8016	0.0017
	5	0.4761	0.0954	0.7995	0.0014
	10	0.4761	0.0994	0.7913	0.0008
	20	0.4761	0.1042	0.7811	0.0004
	30	0.4761	0.1072	0.7748	0.0003
	40	0.4761	0.1094	0.7702	0.0002
	50	0.4761	0.1111	0.7667	0.0002
	60	0.4761	0.1124	0.7639	0.0001
	70	0.4761	0.1136	0.7614	0.0001
	80	0.4761	0.1146	0.7594	0.0001
	90	0.4761	0.1154	0.7575	0.0001
	100	0.4761	0.1162	0.7558	0.0001
Unequal	1	0.4376	0.0816	0.8135	0.0043
	2	0.4376	0.0822	0.8122	0.0030
	3	0.4376	0.0832	0.8098	0.0023
	4	0.4376	0.0844	0.8072	0.0019
	5	0.4376	0.0854	0.8048	0.0016
	10	0.4376	0.0898	0.7948	0.0008
	20	0.4376	0.0950	0.7828	0.0004
	30	0.4376	0.0983	0.7755	0.0003
	40	0.4376	0.1006	0.7702	0.0002
	50	0.4376	0.1023	0.7662	0.0002
	60	0.4376	0.1038	0.7629	0.0002
	70	0.4376	0.1050	0.7601	0.0001
	80	0.4376	0.1060	0.7577	0.0001
	90	0.4376	0.1070	0.7556	0.0001
	100	0.4376	0.1078	0.7537	0.0001

Note. No penalty is assessed if $d = 1$.

Table C2 Estimated Parameters for Penalized Best Linear Predictor for GRE Writing: Assessment Accuracy

Weight	d	$\sigma^2(O)$	$MSE_{M_1 d}$	$\rho^2_{M_1 d}$	$\sigma^2\left(E\left(r_{M_{1d}}\|G\right)\right)$
Equal	1	0.4965	0.0713	0.8564	0.0036
	2	0.4965	0.0720	0.8550	0.0020
	3	0.4965	0.0729	0.8532	0.0013
	4	0.4965	0.0737	0.8516	0.0010
	5	0.4965	0.0743	0.8504	0.0008
	10	0.4965	0.0763	0.8463	0.0005
	20	0.4965	0.0786	0.8418	0.0003
	30	0.4965	0.0802	0.8386	0.0002
	40	0.4965	0.0815	0.8359	0.0002
	50	0.4965	0.0827	0.8335	0.0002
	60	0.4965	0.0838	0.8313	0.0002
	70	0.4965	0.0848	0.8293	0.0001
	80	0.4965	0.0857	0.8274	0.0001
	90	0.4965	0.0866	0.8256	0.0001
	100	0.4965	0.0874	0.8239	0.0001

Table C2 Continued

Weight	d	$\sigma^2(O)$	$\mathrm{MSE}_{\mathrm{M}_1 d}$	$\rho^2_{\mathrm{M}_1 d}$	$\sigma^2\left(E\left(r_{\mathrm{M}_1 d}\vert G\right)\right)$
Unequal	1	0.4961	0.0711	0.8567	0.0037
	2	0.4961	0.0718	0.8554	0.0020
	3	0.4961	0.0727	0.8535	0.0013
	4	0.4961	0.0735	0.8519	0.0010
	5	0.4961	0.0741	0.8506	0.0008
	10	0.4961	0.0762	0.8465	0.0005
	20	0.4961	0.0784	0.8420	0.0003
	30	0.4961	0.0800	0.8388	0.0002
	40	0.4961	0.0813	0.8362	0.0002
	50	0.4961	0.0824	0.8338	0.0002
	60	0.4961	0.0835	0.8317	0.0002
	70	0.4961	0.0845	0.8297	0.0001
	80	0.4961	0.0854	0.8278	0.0001
	90	0.4961	0.0863	0.8261	0.0001
	100	0.4961	0.0871	0.8244	0.0001

Note. No penalty is assessed if $d = 1$.

Table C3 Estimated Parameters for Penalized Best Linear Predictor for Praxis Writing: Assessment Accuracy

d	$\sigma^2(O)$	$\mathrm{MSE}_{\mathrm{M}_1 d}$	$\rho^2_{\mathrm{M}_1 d}$	$\sigma^2\left(E\left(r_{\mathrm{M}_1 d}\vert G\right)\right)$
1	0.2480	0.0992	0.6002	0.0026
2	0.2480	0.0993	0.5997	0.0023
3	0.2480	0.0996	0.5984	0.0022
4	0.2480	0.1001	0.5964	0.0019
5	0.2480	0.1007	0.5941	0.0018
10	0.2480	0.1043	0.5796	0.0012
20	0.2480	0.1136	0.5510	0.0007
30	0.2480	0.1169	0.5287	0.0004
40	0.2480	0.1211	0.5119	0.0003
50	0.2480	0.1243	0.4989	0.0002
60	0.2480	0.1268	0.4886	0.0002
70	0.2480	0.1289	0.4803	0.0001
80	0.2480	0.1306	0.4734	0.0001
90	0.2480	0.1320	0.4677	0.0001
100	0.2480	0.1332	0.4627	0.0001

Note. No penalty is assessed if $d = 1$.

Table C4 Estimated Mean Score Differences (Equal Weights) for Assessment Accuracy: TOEFL Writing

	d														
Subgroup	1	2	3	4	5	10	20	30	40	50	60	70	80	90	100
1 (15,592)	.006	−.014	.027	−.034	−.039	−.049	−.047	−.042	−.038	−.034	−.031	−.028	−.026	.024	.022
2 (4,374)	.044	.028	.018	.012	.007	−.001	−.001	.0008	.003	.005	.006	.007	.008	.009	.010
3 (67,464)	.046	.030	.020	.014	.009	−.004	−.011	−.013	−.014	−.015	−.015	−.015	−.015	−.015	−.015
4(14,615)	.071	.057	.047	.041	.037	.026	.019	.017	.015	.015	.014	.013	.013	.013	.012
5 (47,016)	.019	.016	.014	.013	.012	.008	.005	.003	.002	.001	.0009	.0004	.0001	−.0003	−.0006
6 (11,582)	.031	.026	.023	.020	.019	.014	.011	.009	.007	.006	.006	.005	.005	.004	.004
7 (372,583)	−.019	−.009	−.003	.001	.004	.010	.012	.013	.013	.013	.013	.013	.013	.013	.013
8 (49,625)	.025	.031	.034	.036	.037	.038	.035	.033	.030	.028	.027	.025	.024	.023	.022
9 (3,867)	.038	.028	.022	.018	.016	.011	.010	.010	.010	.011	.011	.011	.011	.011	.011
10 (114,228)	.012	.002	−.004	−.009	−.012	−.020	−.025	−.027	−.028	−.029	−.029	−.029	−.029	−.029	−.029
11 (14,209)	.053	.044	.039	.036	.033	.027	.022	.019	.017	.015	.014	.013	.012	.012	.011
12 (80,157)	−.002	−.017	−.027	−.034	−.039	−.054	−.064	−.067	−.068	−.069	−.069	−.096	−.068	−.068	−.067
13 (9,478)	.092	.079	.071	.065	.060	.045	.033	.026	.022	.019	.016	.015	.013	.012	.011
14 (52,148)	.008	.016	.022	.027	.030	.040	.046	.048	.049	.049	.048	.048	.048	.047	.047
15 (6,482)	.019	.022	.025	.027	.029	.033	.034	.033	.031	.030	.029	.027	.026	.025	.025
16 (69,625)	−.114	−.010	−.086	−.078	−.071	−.052	−.034	−.026	−.021	−.017	−.015	−.013	−.011	−.010	−.009
17 (12,008)	−.041	−.033	−.027	−.023	−.020	−.012	−.007	−.006	−.005	−.005	−.004	−.004	−.004	−.004	−.004
18 (38,027)	.038	.023	.014	.008	.004	−.004	−.004	−.002	−.0006	.0009	.002	.003	.004	.004	.005
19 (23,474)	.137	.120	.109	.101	.096	.083	.076	.073	.071	.070	.068	.067	.066	.065	.064

Note. No penalty is assessed if $d = 1$. Values in parentheses represent the sample sizes for each subgroup.

Table C5 Estimated Mean Score Differences (Unequal Weights) for Assessment Accuracy: TOEFL Writing

	d														
Subgroup	1	2	3	4	5	10	20	30	40	50	60	70	80	90	100
1 (15,592)	−.002	−.025	−.038	−.047	−.053	−.064	−.063	−.059	−.055	−.052	−.049	−.047	−.045	−.043	−.041
2 (4,374)	.037	.018	.007	−.0002	−.005	−.015	−.017	−.015	−.013	−.012	−.011	−.010	−.009	−.009	−.008
3 (67,464)	.041	.024	.013	.005	.0006	−.014	−.022	−.024	−.026	−.026	−.027	−.027	−.027	−.027	−.027
4 (14,615)	.065	.049	.038	.031	.026	.014	.006	.003	.002	.001	.0003	−.0003	−.0007	−.001	−.001
5 (47,016)	.019	.016	.014	.012	.011	.007	.003	.001	−.0004	−.001	−.002	−.003	−.004	−.005	−.005
6 (11,582)	.029	.023	.019	.016	.014	.009	.005	.002	.0006	−.001	−.002	−.002	−.003	−.004	−.004
7 (372,583)	−.016	−.006	.0007	.005	.008	.015	.019	.020	.020	.020	.020	.020	.020	.020	.020
8 (49,625)	.026	.032	.036	.038	.039	.041	.039	.036	.034	.033	.031	.030	.029	.028	.027
9 (3,867)	.031	.020	.013	.008	.006	−.0002	−.002	−.002	−.002	−.002	−.002	−.002	−.002	−.002	−.002
10 (114,228)	.008	−.004	−.011	−.016	−.019	−.028	−.034	−.037	−.038	−.039	−.039	−.039	−.040	−.040	−.040
11 (14,209)	.048	.039	.033	.029	.026	.019	.013	.009	.007	.006	.004	.003	.002	.001	.001
12 (80,157)	.002	−.014	−.025	−.032	−.038	−.053	−.063	−.066	−.068	−.068	−.069	−.069	−.068	−.068	−.067
13 (9,478)	.098	.085	.076	.070	.064	.049	.037	.030	.026	.023	.021	.019	.017	.016	.015
14 (52,148)	.005	.013	.019	.023	.026	.035	.041	.042	.043	.042	.042	.041	.041	.040	.040
15 (6,482)	.016	.020	.023	.025	.026	.029	.030	.029	.027	.025	.024	.023	.022	.021	.020
16 (69,625)	−.108	−.090	−.078	−.069	−.062	−.042	−.024	−.015	−.009	−.005	−.002	−.0001	.002	.003	.004
17 (12,008)	−.038	−.029	−.023	−.019	−.016	−.008	−.003	−.001	−.0004	−.0000	.0002	.0004	.0006	.001	.001
18 (38,027)	.036	.019	.009	.002	−.003	−.012	−.013	−.011	−.009	−.007	−.006	−.005	−.004	−.003	−.003
19 (23,474)	.136	.117	.105	.097	.091	.077	.069	.067	.065	.064	.063	.062	.061	.061	.060

Note. No penalty is assessed if $d = 1$. Values in parentheses represent the sample sizes for each subgroup.

Table C6 Estimated Mean Score Differences (Equal Weights) for Assessment Accuracy: GRE Writing

	d														
Subgroup	1	2	3	4	5	10	20	30	40	50	60	70	80	90	100
1 (7,431)	.040	.043	.044	.044	.044	.043	.040	.038	.036	.034	.033	.031	.030	.029	.029
2 (4,374)	.012	.021	.025	.027	.028	.028	.025	.021	.018	.016	.014	.013	.012	.011	.010
3 (1,808)	−.005	−.003	−.002	−.002	−.002	−.003	−.006	−.008	−.009	−.010	−.011	−.012	−.013	−.013	−.014
4 (18,807)	−.003	.008	.014	.017	.019	.024	.025	.025	.024	.023	.023	.022	.022	.021	.021
5 (1,223)	.003	.008	.010	.012	.012	.012	.008	.005	.003	.001	−.0001	−.001	−.002	−.003	−.003
6 (3,265)	.0004	.001	.001	.0009	.0008	−.0002	−.002	−.004	−.006	−.007	−.008	−.009	−.009	−.01	−.011
7 (31,677)	.041	.052	.057	.059	.060	.059	.053	.048	.045	.042	.040	.039	.037	.036	.035
8 (2,808)	.074	.069	.067	.065	.065	.063	.062	.061	.061	.059	.058	.057	.056	.054	.053
9 (1,264)	.023	.027	.028	.029	.029	.028	.024	.020	.018	.015	.014	.012	.011	.009	.008
10 (2,159)	.065	.075	.079	.081	.082	.082	.078	.073	.070	.067	.064	.062	.060	.059	.057
11 (5,025)	−.027	−.030	−.031	−.032	−.032	−.033	−.032	−.032	−.031	−.030	−.030	−.029	−.028	−.028	−.028
12 (6,846)	−.040	−.053	−.058	−.061	−.062	−.061	−.054	−.048	−.043	−.039	−.036	−.033	−.031	−.029	−.027
13 (6,669)	−.036	−.044	−.047	−.049	−.049	−.049	−.045	−.041	−.039	−.036	−.034	−.033	−.031	−.030	−.029
14 (725)	−.034	−.042	−.046	−.047	−.048	−.049	−.046	−.043	−.040	−.038	−.037	−.035	−.034	−.033	−.032
15 (2,888)	−.014	−.020	−.022	−.024	−.024	−.025	−.024	−.022	−.021	−.020	−.020	−.019	−.019	−.018	−.018
16 (45,447)	.005	.001	−.0003	−.001	−.002	−.002	−.001	−.0005	−.0005	.0003	.0005	.0007	.0008	.0009	.0009
17 (52,435)	−.028	−.034	−.037	−.038	−.039	−.039	−.037	−.035	−.033	−.032	−.030	−.029	−.029	−.028	−.027

Note. No penalty is assessed if $d = 1$. Values in parentheses represent the sample sizes for each subgroup.

Table C7 Estimated Mean Score Differences (Unequal Weights) for Assessment Accuracy: GRE Writing

	d														
Subgroup	1	2	3	4	5	10	20	30	40	50	60	70	80	90	100
1 (7,431)	.040	.043	.044	.044	.044	.043	.040	.038	.036	.034	.033	.031	.030	.029	.029
2 (4,374)	.012	.021	.025	.027	.029	.025	.026	.022	.020	.017	.016	.014	.013	.012	.012
3 (1,808)	−.004	−.003	−.002	−.002	−.002	−.003	−.005	−.007	−.009	−.010	−.011	−.012	−.012	−.013	−.013
4 (18,807)	−.003	.008	.014	.017	.020	.025	.026	.026	.025	.025	.024	.023	.023	.022	.022
5 (1,223)	.003	.008	.011	.012	.012	.012	.009	.006	.004	.002	.0008	−.0003	−.001	−.002	−.002
6 (3,265)	.001	.001	.001	.001	.001	−.0001	−.002	−.004	−.005	−.007	−.008	−.008	−.009	−.010	−.010
7 (31,677)	.041	.052	.057	.059	.060	.059	.053	.048	.045	.042	.040	.039	.037	.036	.035
8 (2,808)	.074	.069	.067	.065	.064	.063	.062	.062	.061	.059	.058	.057	.056	.054	.053
9 (1,264)	.023	.027	.028	.029	.029	.028	.024	.021	.018	.016	.014	.012	.011	.010	.009
10 (2,159)	.065	.075	.079	.081	.083	.083	.078	.074	.071	.068	.065	.063	.061	.059	.058
11 (5,025)	−.027	−.030	−.031	−.032	−.032	−.033	−.032	−.032	−.031	−.030	−.030	−.029	−.029	−.028	−.028
12 (6,846)	−.041	−.053	−.058	−.061	−.062	−.062	−.055	−.049	−.044	−.040	−.037	−.034	−.032	−.030	−.028
13 (6,669)	−.036	−.044	−.047	−.049	−.050	−.050	−.046	−.042	−.039	−.037	−.035	−.033	−.032	−.031	−.030
14 (725)	−.034	−.042	−.046	−.048	−.049	−.049	−.046	−.043	−.041	−.039	−.037	−.036	−.035	−.033	−.032
15 (2,888)	−.014	−.020	−.022	−.024	−.025	−.025	−.024	−.023	−.022	−.021	−.020	−.020	−.019	−.019	−.018
16 (45,447)	.005	.001	−.0003	−.001	−.002	−.002	−.001	−.0006	−.0001	.0002	.0004	.0006	.0007	.0008	.0008
17 (52,435)	−.028	−.034	−.037	−.038	−.039	−.039	−.037	−.035	−.033	−.032	−.031	−.030	−.029	−.028	−.028

Note. No penalty is assessed if $d = 1$. Values in parentheses represent the sample sizes for each subgroup.

Table C8 Estimated Mean Score Differences for Assessment Accuracy: Praxis Writing

	d														
Subgroup	1	2	3	4	5	10	20	30	40	50	60	70	80	90	100
1 (18,362)	.034	.039	.043	.047	.050	.060	.070	.074	.077	.079	.080	.081	.081	.082	.082
2 (4,583)	−.062	−.055	−.049	−.044	−.039	−.022	−.004	.006	.012	.016	.020	−.022	.025	0.026	.028
3 (7,822)	.010	.014	.017	0.019	.021	.029	.037	.041	.043	.044	.045	.046	.047	.047	.047
4 (3,583)	−.002	−.001	−.0005	.0003	.0009	.003	.006	.007	.007	.007	.008	.008	.008	.008	.008
5 (944)	.005	.005	.006	.006	.006	.007	.008	.008	.009	.009	.009	.010	.010	.010	.011
6 (9,599)	.002	.004	.005	.006	.008	.012	.016	.018	.019	.020	.020	.020	.020	.020	.021
7 (104,820)	−.004	−.006	−.007	−.008	−.009	−.013	−.016	−.018	−.020	−.020	−.020	−.021	−.021	−.021	−.021

Note. No penalty is assessed if $d = 1$. Values in parentheses represent the sample sizes for each subgroup.

Suggested citation:

Yao, L., Haberman, S. J., & Zhang, M. (2019). *Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors* (Research Report No. RR-19-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12248

Action Editor: Rebecca Zwick

Reviewers: Hongwen Guo